TOEFL®

Research Reports

RR - 78
November 2004

Toward Accessible Computer-Based Tests: Prototypes for Visual and Other Disabilities

Eric G. Hansen

Douglas C. Forer

Moon J. Lee

Toward Accessible Computer-Based Tests:
Prototypes for Visual and Other Disabilities

Eric G. Hansen and Douglas C. Forer
ETS, Princeton, NJ

Moon J. Lee
Washington State University

RR-04-25

ISSN 1552-8219

Abstract

There is a great need to explore approaches for developing computer-based testing systems that are more accessible for people with disabilities. This report explores three prototype test delivery approaches, describing their development and formative evaluations. Fifteen adults, 2 to 4 from each of the six disability statuses—blindness, low vision, deafness, deaf-blindness, learning disability, and no disability—participated in a formative evaluation of the systems. Each participant was administered from 2 to 15 items in each of one or two of the systems. The study found that although all systems had weaknesses that should be addressed, almost all of the participants (13 of 15) would recommend at least one of the delivery methods for high-stakes tests, such as those for college or graduate admissions. The report concludes with recommendations for additional research that testing organizations seeking to develop accessible computer-based testing systems can consider.

Key words: Accessibility, computer-based testing, disabilities

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service® (ETS®) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

❖ ❖ ❖

A continuing program of research related to the TOEFL test is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, reviews and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Because the studies are specific to the TOEFL test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. Many projects require the cooperation of other institutions, however, particularly those with programs in the teaching of English as a foreign or second language and applied linguistics. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (2004-2005) members of the TOEFL Committee of Examiners are:

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org

Web site: www.ets.org/toefl

Abstract

There is a great need to explore approaches for developing computer-based testing systems that are more accessible for people with disabilities. This report explores three prototype test delivery approaches, describing their development and formative evaluations. Fifteen adults, 2 to 4 from each of the six disability statuses—blindness, low vision, deafness, deaf-blindness, learning disability, and no disability—participated in a formative evaluation of the systems. Each participant was administered from 2 to 15 items in each of one or two of the systems. The study found that although all systems had weaknesses that should be addressed, almost all of the participants (13 of 15) would recommend at least one of the delivery methods for high-stakes tests, such as those for college or graduate admissions. The report concludes with recommendations for additional research that testing organizations seeking to develop accessible computer-based testing systems can consider.

Key words: Accessibility, computer-based testing, disabilities

Acknowledgments

The authors acknowledge the Test of English as a Foreign Language™ (TOEFL®) program, the Graduate Record Examinations® (GRE®) program, and the ETS Research Area for their generous financial support; Kevin Bentley, Rob Rarich, and Mike Wagner for system design and development; Ruth Loew for helpful input at many points and participation in deafness-related evaluations; Beth Brownstein for assistance with item identification and selection; Carol Myford for input on instrument design; René Lawless for recruiting; Amanda McBride and Kim Fryer for editing; Amy Chen for data preparation and assistance with field evaluations; Lisa Hemat for recruiting and materials preparation; Janet Stumper for graphics preparation; and Maria Rossi for administrative support. Also providing helpful input were Linda Cook, David Denotaris, Crista Earl, Daniel Eignor, Ramin Hemat, Jay Leventhal, Robert Mislevy, Judith Mounty, Charlene Sorensen, Matthew Tedder, Clarence H. Thomas, and several anonymous reviewers. We also would like to thank the following for recruiting and for coordinating evaluations: Laurie McArthur, Jutta Treviranus, and staff of the Adaptive Technology Resource Centre (ATRC) at the University of Toronto, Ontario, Canada; Robert Weinstock and staff at Gallaudet University, Washington, D.C.; Joseph McNulty, Susan Ruzenski, and staff at the Helen Keller National Center for Deaf-Blind Youths and Adults, Sands Point, New York; Cheryl Stewart and staff at Overbrook School for the Blind, Philadelphia, Pennsylvania; and Crista Earl and staff at the American Foundation for the Blind, New York, New York. The views expressed are solely those of the authors.

Table of Contents

Page

Appendixes

List of Tables

Part 1: Introduction

There is a great need to ensure that computer-based tests are as accessible as possible for people with disabilities. ETS has considerable experience in administering tests to individuals with disabilities. ETS programs often provide large-print, hard-copy Braille, and audiocassette versions of tests, and human readers, for individuals with disabilities. However, some disability-advocacy groups and related professional organizations have expressed concern that ETS *computer-based* tests are not compatible with assistive technology (sometimes called “adaptive” technology) used by individuals who are blind or have other visual disabilities (American Council of the Blind, 2000). While ETS computer-based tests do support some assistive technologies, such as screen magnification software (ZoomText), alternative keyboard, and alternative pointing devices, they do not currently support the use of screen readers[1] or refreshable Braille displays[2] (technologies explained later in this report).

Some Trends

In considering how to make computer-based tests more accessible, one needs to take into account some recent trends.

Improvements in Technology

Computer technologies and related technologies, including assistive technologies such as screen readers and refreshable Braille displays, have improved dramatically over the past few years. These technologies are becoming an integral part of how individuals with disabilities interact with the world. Individuals with disabilities expect to use such technologies when they interact with computers. While the growing availability of technology increases individual and societal demands for technology in testing settings, it also provides means for fulfilling those demands.

Societal Demands for Inclusion

There are increasing demands in our society to ensure that products and services are accessible to people with disabilities. Among organizations developing computer-based material of any kind, and computer-based tests in particular, many are giving closer attention to the concept of inclusive design strategies (also termed *universal design* strategies) through which a software product might be made more usable by everyone (Heath & Hansen, 2002; IMS Global

Learning Consortium, 2002; Thompson, Johnstone, & Thurlow, 2002; Thompson, Thurlow, Quenemoen, & Lehr, 2002). By coupling inclusive design with design for "re-use" of content, it seems possible to build integrated systems that can be used by individuals from diverse disability groups as well as by individuals without any disability. Such systems might achieve that integration through internal features (direct access) or by compatibility with assistive technology (compatible accessibility), or both.

The Need for a More Robust Framework for Understanding Validity

In building accessible computer-based tests, testing organizations have obligations to understand and investigate complex issues bearing on the validity of interpretations arising from use of such tests (Bennett, 1999a). There is a growing recognition that effort to develop accessible computer-based testing must be instituted within some larger and more robust validity framework in order to minimize the possibility that alterations[3] invalidate test results (Hansen, Forer, & Lee, 2001; Heath & Hansen, 2002).

There is a need to take these trends into account and to explore approaches to developing computer-based testing systems that would be more inclusive yet that would preserve or enhance the validity of inferences arising from tests administered to individuals with disabilities. Such exploration should deepen our practical understanding of issues of inclusion, validity, and technology as they relate to computer-based testing. Such exploration should pay attention to strategies for meeting diverse access requirements more efficiently.

Purpose

This report explores three test delivery approaches, describing their development and formative evaluations. Each approach is implemented as a prototype system that integrates a number of accessibility features that are intended for individuals of one or more disability statuses. Fifteen adults, 2 to 4 from each of six disability statuses—blindness (BL), low vision (LV), deafness (DF), deaf-blindness (DB), learning disability (LD), and no disability (ND)—participated in final (Phase 3) formative evaluation of the systems. Each participant was administered from 2 to 15 items in one or two of the systems. Items came from domains of reading comprehension, listening comprehension, structure (grammar), writing, and math. The participants gave feedback through interviews and focus groups. The key question was: "What are the strengths and weaknesses of the systems in meeting the access requirements of individuals with the six different disability statuses?"

The three systems were:

1. *The Self-Voicing Test* (SVT) provided built-in text-to-speech capabilities and keyboard operation. The intent of this system was to be the most useful for people with visual disabilities (including blindness) and for some with learning disabilities. For individuals who are blind, SVT was meant to provide *direct* accessibility.

2. *The HTML-Form System* (HFS) used standard HTML form input elements (text input boxes, radio buttons, drop-down boxes, check boxes) and supported the optional use of (a) text-to-speech (via screen reader software) and (b) Braille (via refreshable Braille display). Expected users of this system were individuals with a range of disability statuses including individuals who rely on Braille (i.e., individuals who are blind or deaf-blind). For individuals who are blind, HFS was meant to be *compatibly* accessible with screen reader technology.

3. *The Visually Oriented System* (VOS) was visually oriented—similar to current ETS computer-based tests—and operable via mouse, and was intended for individuals who are deaf or nondisabled. The VOS was originally intended to be compatibly accessible with screen reader technology and therefore useful for individuals who are blind, but an early evaluation showed that it was not fulfilling this requirement, so in the last phase of the effort, its use was focused on others (i.e., individuals who are deaf).

Emphases of This Report

While this report deals with the broad issues of *technology*, *inclusion*, and *validity* as they relate to computer-based testing, there are special emphases or foci. For example, while the report focuses on inclusion of individuals with diverse disabilities, it gives special focus to the areas of blindness and other visual disabilities. This extra attention to blindness and other visual disabilities is appropriate since most educational content, including educational tests, is highly *visual* in nature, which places individuals with visual disabilities (e.g., blindness, deaf-blindness, and low vision) at a particular disadvantage. Furthermore, while the report deals with some of the *validity* issues, the greatest emphasis for the central part of the report is on the *technologies* and the usability of their interfaces to the users with diverse disabilities.

Parts of This Report

Following this part (Part 1) of the document, Part 2 gives background on key concepts and technologies, and describes the development of the three systems. It begins by providing more background on assistive technologies (e.g., screen readers and refreshable Braille displays).

Part 3 describes the method for the Phase 3 formative evaluation. Part 4 describes the results for the Phase 3 formative evaluation. Organized in a question-and-answer format, Part 5 describes selected highlights of results of the research. Part 6 provides discussion and conclusions, including recommendations for organizations desiring to improve the accessibility of computer-based tests.

Benefits of This Kind of Study

Formative research studies of this kind in the area of disability access are important for several reasons. First, they illuminate the general "topography" of disability access. That is, they provide an overview of how individuals with diverse disabilities interact with different technological solutions. Second, they provide clues about how to implement certain capabilities given the current state of technology. For example, speech synthesis could be provided by a self-voicing approach or by a screen reader. This kind of study lets one learn about the challenges of implementing such capabilities in either approach. Third, they may illuminate the benefits and challenges that might come from efforts that would aggregate still more capabilities into fewer systems or even a single integrated system.

Part 2: Background and Development of the Systems

This part provides background on key concepts and technologies, and describes the development of systems (including their early formative evaluations).

Background Concepts, History, and Rationale

This section elaborates on concepts in the introduction for readers unfamiliar with them.

The Need for Improvements in Accessibility

Society expects that computer-based products and services available to the public in general will be accessible to individuals with disabilities. ETS has long administered tests to individuals with disabilities in ways that can remove accessibility barriers. ETS programs often provide large-print, hard-copy Braille, and audiocassette versions of tests. Optional supplements may include large-print diagrams and raised line Braille diagrams.[4] Human readers may read content to the test taker and may write down the test taker's answers.[5] Yet, some disability-advocacy groups and related professional organizations have expressed concern that ETS *computer-based* tests are not compatible with assistive technology (sometimes called "adaptive" technology) used by individuals who are blind or have other visual disabilities (American Council of the Blind, 2000).

In the context of access to software, assistive technology includes products such as the following:[6]

- Screen magnifiers: used by people with visual disabilities to enlarge and change colors on the screen to improve the visual readability of text and images.
- Screen readers: used by people who are blind or have print-related disabilities to read text via synthesized speech or Braille displays.
- Voice recognition software: used by people who have particular physical disabilities, especially disabilities that limit the mobility of hands and arms.
- Alternative keyboards: used by people with certain physical disabilities to simulate or replace the standard keyboard.
- Alternative pointing devices: used by people with certain physical disabilities to simulate mouse pointing and button activation (Jacobs, Gunderson, & Hansen, 2002).

While ETS does offer assistive technologies such as screen magnification software (ZoomText), alternative keyboard, and alternative pointing devices, screen readers are not among the alterations for computer-based testing listed by ETS on its Web site (ETS, 2002).[7]

Screen Reader Technology

What is a screen reader? Screen reader programs are commercial software packages that allow a person who is blind (or has certain other specific disabilities) to interact with a variety of other software—such as word processors, spreadsheets, and Web browsers—using text-to-speech technology.[8] Using a screen reader, the user can hear the content being worked with (e.g., the Web content or word processing text content) as well as hear information about the controls on the application (e.g., menus, buttons, dialogs). The more capable screen reader packages, such as JAWS for Windows (or JFW, Freedom Scientific) or Window-Eyes (GW Micro), provide a very rich set of functionalities. One can choose among a variety of speech characteristics (rate, gender, voice, etc.) and navigate through passages of paragraphed text (word-by-word, character-by-character, etc.), as well as through more complex content, such as tables, forms, and frames.

Refreshable Braille Displays

Some screen reader packages support refreshable Braille displays. A refreshable Braille display has hardware that raises and lowers Braille dot patterns on command from a computer. Users read the Braille dot patterns by moving their fingertips across the Braille panel, much as they would read the Braille dot patterns on paper. Unlike the static medium of paper, a refreshable Braille display is dynamic; it "refreshes." The dot patterns change under control of the user to display new content. The "dots" are actually the ends of pins that are mechanically moved up or down, as appropriate, through holes in a flat surface. When a user has read the single row of Braille characters representing a portion of the target content, the user can then navigate to a different location in content causing the dot patterns to change. A Braille display usually presents no more than about 40 Braille characters at time, though some have capacities of 80 characters or more. If the number of characters sent to the Braille display exceeds the size of display, the display stores the extra characters in memory where the user can retrieve them. Users control the Braille display through keyboard commands to the computer (to which it is attached) and/or through operable controls on the display hardware itself. Most refreshable Braille displays for use with desktop computers cost several thousand dollars or more. Refreshable Braille displays are becoming increasingly popular as a feature on electronic note taker devices that provide keyboard input and synthesized speech output. When operated under control of screen reader software, the Braille display can present the same (or similar) information that a hearing person hears via synthesized speech. Refreshable Braille support benefits Braille users who are not only blind but who are deaf-blind and unable to hear the synthesized speech.

Limitations of Early Screen Readers

Early screen readers had some severe limitations. The first screen reader programs lacked support for refreshable Braille. Furthermore, until about three years ago, the internal Microsoft Windows screen model was not sophisticated enough for representing (a) the operating environment (the Program Manager, multiple windows, icons, etc.), (b) the application (e.g., word processor, spreadsheet program, Web browser), and (c) the document (word processor document, spreadsheet document, Web page, etc.) in a way that could be readily communicated to a print-disabled user. This lack of a sophisticated screen model—and supporting application programming interface (API)—within Windows placed a significant burden on developers of screen readers to develop their own screen model and then to employ that model to communicate to the person with a print-related disability. This typically meant that screen readers could only read pages left to right and top to bottom. In a Web page with two side-by-side columns, for example, the screen reader might speak the first line of the first column, followed by the first line of the second column and then the second line of the first column making multicolumn content (including tables) and forms (e.g., filling in addresses) difficult to handle.

A major advance occurred with the advent of Microsoft Active Accessibility (MSAA), which provided a more sophisticated screen model. MSAA provides software "hooks" to work compatibly with screen readers, screen magnifiers, or alternative input devices such as adapted keyboards or single switches. Microsoft Internet Explorer 5.0, which was released in about 1999, was the first Web browser to implement MSAA, and since then vendors of screen readers have developed their software to exploit this more sophisticated model. Handling of difficult content, such as frames, forms, and multicolumn content, has improved greatly as has support for refreshable Braille. The more advanced and popular screen readers, such as JFW or Window-Eyes, provide a very rich set of functionalities that exploit MSAA. The role played by MSAA is similar to that played by the Java Accessibility API developed by Sun Microsystems.

Notwithstanding these advances, a screen reader with or without a Braille display can be challenging to use for several reasons. First, it can be difficult to learn how to use screen reader software. Some of this difficulty is due to the sheer richness of the functionalities available. For example, JAWS has five different navigation modes ("cursor modes") that are suited to different kinds of content and purpose, such as reading content, filling out forms, or using Web pages or some other kind of content.[9] This means that there may be more than one way to access the same content, and considerable expertise may be necessary to know the optimal mode for a given situation. A second issue concerns the content design. A Web page might be readily accessible using JFW but much less so using Window-Eyes or vice versa. Even within a single Web page accessed twice by the same screen reader, the result may be somewhat different for the user in each instance, depending on changes in the system state between accesses (C. Earl, personal communication, October 31, 2001). Some kinds of content structures are known to be more difficult to use, such as “forms” and “frames.”[10] Third, screen readers often tell the user more information than is helpful or sometimes less information than is necessary. While popular screen reader programs often allow the user to control the verbosity of the screen reader (how much it says about what it knows), this may provide small comfort to test administrators who want to ensure that all test takers have access to the same important information. Fourth, the quality of the output from the screen reader can depend on the viewer (browser or other application) with which it is being used. For example, at this time, the Internet Explorer Web browser provides better support for the most popular screen readers than does the Netscape Communicator Web browser, which does not currently support MSAA. One result of these challenges is that when something goes wrong, it is often difficult to identify whether the problem is due to the authored content, the type of viewer (browser) used, the user’s level of screen reader or Braille display expertise, or hardware issues.

The challenges of accessing the Web via screen reader are underscored by a study which found that sighted participants (using no assistive technology) were about six times as successful in completing tasks as people using screen readers (Coyne & Nielsen, 2001, page 5).

Such challenges, discussed later in greater detail, can be partially mitigated by authoring test content in ways that do not demand high levels of expertise and that minimize the impact of differences between Web user agents (browsers). One can also standardize on a particular brand and version of browser. Yet, as one considers the benefits and challenges of using screen readers, the question arises about possible alternative ways of obtaining access to computer-based tests.

Self-Voicing Applications

In contrast to the screen reader approach, *self-voicing applications* provide *built-in* speech output capabilities rather than relying on *external* assistive technology, such as *screen reader* software.[11] The user of a self-voicing application does not need to start another software application (i.e., a screen reader) to hear and navigate through the content. In principle, a self-voicing application can have a simplified, easier-to-learn interface, since the speech and navigation capabilities are designed with the application itself. The user of a self-voicing test delivery application need not have more commands than those necessary for test delivery. A self-voicing application has the distinct advantage over the screen reader approach in that the developer of the self-voicing application (e.g., the testing organization) has a very high degree of control over the experience of the test taker. This high level of control allows the testing organization to better ensure consistency of experience across test takers. A possible limitation of the self-voicing approach is that the navigation scheme for a self-voicing test may be unlike that scheme for any other application and would have to be learned.

Directly Accessible Versus Compatibly Accessible

The distinction between screen reader technology and self-voicing applications is representative of a recognized dichotomy in solutions to disability access challenges. Specifically, solutions for making software accessible are sometimes grouped into two categories: direct access and compatible access (IMS Global Learning Consortium, 2002; Rothberg & Wlodkowski, 2003). A product that is directly accessible, such as a self-voicing application, allows a person with a disability to operate all on-screen controls and to access content without relying on an additional software layer—that of the assistive technology. On the other hand, software or content that is compatibly accessible is designed to work with assistive technology. As discussed later, some solutions do not fit neatly into one category.

The Validity of Inferences Arising From Altered Tests

Regardless of the technological solution provided, it is critical that the inferences arising from computer-based tests for people with disabilities be as valid and fair as possible. There is a growing recognition that efforts to lower accessibility barriers in computer-based testing should be instituted within some larger validity framework so that alterations do not invalidate test results (Hansen, Forer, & Lee, 2001; Heath & Hansen, 2002). According to Heath and Hansen (2002):

> Thus, for any given test and test-taker, not every accessibility feature necessarily promotes validity. Accessibility features need to be examined for their impact on validity. Another example concerns the use of a readaloud accommodation[12] in a test of reading comprehension. The test performance of a person who is blind might benefit from having content read aloud to them by synthesized speech or by prerecorded or live human speech. However, suppose that this test of reading comprehension includes decoding words from letters as part of the construct (intent of measurement). In this case, a readaloud accommodation would essentially prevent student performance from serving as a source of evidence about decoding ability since hearing the words spoken would not require decoding in order to answer questions correctly.

This scenario would suggest that providing the readaloud accommodation may invalidate the test and that, therefore, the readaloud accommodation should *not* be allowed. However, the decision about whether to actually allow a readaloud accommodation might involve additional considerations.

Thus, the general issue of improving disability access to computer-based tests must be considered within a framework that recognizes validity as a preeminent technical consideration. Evidence-centered assessment design (ECD), which frames an assessment as embodying an evidentiary argument (Mislevy, Steinberg, & Almond, 2003), has been suggested as a promising approach in this regard (Hansen, Forer, & Lee, 2001; Heath & Hansen, 2002; Hansen, Mislevy, & Steinberg, 2003).

The Need to Identify Leverage Points for Improved Efficiency

To make computer-based tests more accessible, it is important to identify leverage points for improved *efficiency* in meeting diverse access requirements. One general strategy for improving efficiency is "reuse"—using what is already available rather than reinventing or building "from scratch." There is a need to learn more about what kinds of reuse are feasible in the context of testing individuals with disabilities. One strategy for increasing reuse is to take the set of partially overlapping functionalities available in several systems and to integrate them into a single system with the full—but nonoverlapping—set of functionalities

Before building such an integrated system, it is important to ask how usable are the features that are believed to be most fundamental for addressing diverse access needs. What are the strengths and weaknesses of current ways of implementing such features? Is it possible to reuse content or other resources across delivery systems?

Even though the goal may be to have important features included within a single integrated computer-based testing system, it may make sense to study them within multiple systems. For example, any important basic feature may have more than one possible approach, and developing different systems tailored to different approaches allows us to compare them. Furthermore, there may be technological barriers that currently hinder integrating multiple approaches into single systems. Developing several systems that are focused on somewhat different disabled populations can provide rich opportunities for learning about the strengths and weaknesses of their features. In doing so, we can learn the basic "topography" of disability access as it exists in today's technological environment and can begin to see what steps might be necessary to move toward greater integration and efficiency.

Early System Development and Evaluation

Self-Voicing Test—Versions 1 and 2

The first Self-Voicing Test (SVT version 1 [SVT1]) had items from two content areas—reading comprehension and listening comprehension. The reading comprehension content consisted of a single passage and five questions.[13] The listening comprehension test content consisted of five questions—two prerecorded audio stimuli with one question each, and one such stimulus with three questions. All listening questions were from the "TOEFL Sampler" practice materials (ETS, 1999).

The system used the Microsoft Speech Software Development Kit 4.0 and was run in a Web-based environment (Internet Explorer 5.0) on a laptop with the Microsoft Windows NT 4.0 operating system. The laptop was a Compaq Armada 7730MT with a Pentium CPU running at 166 MHz with 64 MB RAM. Microsoft Media Player software was used to play audio-visual stimuli, which were in Apple QuickTime format. Also provided were a full-sized keyboard and external speakers.

The system's key features were as follows.

1. The system used the self-voicing approach—rather than the screen reader approach—for providing speech access to software applications. Participants were free to go back and forth between the question (stem), choices (options), and passage (where applicable).[14]

2. The system provided keyboard access to the various commands.

3. The system provided auditory descriptions for visuals (Chisholm, Vanderheiden, & Jacobs, 1999). These auditory descriptions consisted of a prerecorded audio of a narrator describing the visual. Without auditory descriptions, the content provided uniquely by the visuals would be unavailable to individuals who are blind.

4. The visual display of the system used large white text on a black background, a combination that is often helpful to individuals who are partially sighted.

5. No specific time limit was placed on participants to complete the test.[15] (This feature is an incidental rather than essential part of the approach.)

6. Test takers could not go back to previous items. (This feature is an incidental rather than an essential part of the self-voicing approach. Commonly found in computer-adaptive tests, it is in relatively few linear tests.)

SVT1 was evaluated with 17 individuals with visual disabilities (blindness and low vision) (Hansen, Lee, & Forer, 2002). Researchers conducted interviews, observed the individuals using the system, and held a focus group. Most participants indicated they would "highly recommend" this kind of system, and several who had used human readers in previous testing situations indicated that they would prefer this approach to a human reader. Areas for improvement included speech quality, system responsiveness, navigation, and preparation materials.

The study of version 2 of the Self-Voicing Test (SVT2) focused on individuals with blindness and low vision as well as with individuals with learning disabilities. The more important enhancements for SVT2 included: a) more content areas (including "structure" [or grammar, from TOEFL] and mathematics [GRE], in addition to reading comprehension and listening comprehension, which had been found in SVT1),[16] (b) a larger number of items (15 items instead of the 10 in SVT1), (c) better directions for the test, sections, and items, (d) a faster computer (700 MHz CPU instead of a 166 MHz),[17] and (e) a simplified response interface (i.e., single-selection multiple-choice response format for items that did not originally have that interface).

Several supplements were added to ease the reception of math content.

1. *Text Descriptions of Math Visuals*. Text descriptions of the math figures and tables were provided. These were located immediately below the figure or table as part of the passage (stimulus) and were voiced as part of the passage content. The intention was to fulfill the accessibility requirement to "Provide a text equivalent for every non-text element" (Chisholm, Vanderheiden, & Jacobs, 1999). Such a text equivalent can be presented—as it was in SVT2—both visually and by synthesized speech.[18]

2. *Braille and Large-Print Figure Supplements*. Braille and large-print figure supplements were used as part of the system, although not with complete consistency.

The work with SVT2 obliged the developers to confront the issues that some innovative response formats can be difficult both to develop and to use for some students with disabilities.

Each new response format requires its own familiarization materials. Each new command can increase the time needed to learn to use the system and may increase the possibility for forgetting a command or confusing one command with another. To minimize unnecessary multiplicity of formats, for SVT2, all items that *did not already* use an ordinary single-selection multiple-choice format were *adapted* to that format. Such adaptations were made for three items, as described in Table 1. This change was expected to minimize the difficulty of learning how to use the system.

Table 1

Adaptation to Single-selection Multiple-choice Format for Three Items for SVT2

Content type	Response type	
	Original computer-based test implementation	SVT2 – Adaptation to single-selection multiple-choice response format
Example Item 1: Reading Comprehension	Insert Text. Click on one of five location markers in the passage to indicate where the target sentence should be placed.	Select the best location for the target sentence from a list of five locations: “After the first sentence,” “After the second sentence,” and so on to the fifth sentence.
Example Item 2: Structure – Sentence Correction	Click on the Word. Click on the word in the sentence that needs to be corrected in order for the sentence to make sense. Four words are underlined in the sentence.	Select the word that needs to be corrected from a list of words. Test takers use the Control-Down arrow and the Control-Up arrow to navigate between choices, and press Enter to select a choice.
Example Item 3: Listening Comprehension	Multiple-Selection Multiple Choice. Click on two correct answers from a list of choices.	Select one answer on each of two single-selection multiple-choice items on separate pages.

SVT2 was evaluated with a convenience sample of 20 individuals with these disability statuses: BL (*n* = 6), LV (*n* = 7), or LD (*n* = 7). Of the 20 participants, 14 found the synthesized speech helpful, and 17 found it easy to learn the key combinations and other keyboard commands for navigation and control. Furthermore, 7 of the 13 who had previously used a human reader for a test preferred the SVT2 approach to using a human reader. Running the system on a faster computer appeared to substantially address the problem noted in the study of SVT1 of lack of system responsiveness. The areas identified for improvement in SVT2 included the need for repeatable directions, mouse support, strategies for reducing the possibility that a test taker would miss essential content, improved speech quality, better familiarization materials, and improved system robustness.

The formative evaluations of the SVT1 and SVT2 systems showed generally positive user reactions. For example, some users of human readers said they would prefer the self-voicing test approach to a human reader, citing among other advantages the possibility of taking tests *independently*. Yet, in addition to a variety of usability problems, the self-voicing test did not support refreshable Braille, an important feature for many individuals who are blind, and even more critical for individuals who are deaf-blind. Nor did it support constructed responses. Screen reader technology has improved greatly over the past several years and provides support for both refreshable Braille and for user entry of text, which is essential for constructed response formats. Given the improvements in screen reader and Web browser technologies, might current screen reader technology coupled with a nonvoicing but screen-reader-friendly testing application meet the requirements of *nonsighted* users (blind, deaf-blind, and some with low vision)?[19] Furthermore, could this nonvoicing test be made effective and visually pleasing for *sighted* users, including individuals who are deaf (but not deaf-blind), learning disabled, or nondisabled? If so, the screen reader approach might be more attractive now than it was when the self-voicing test effort was initiated. In response to this question, a visually oriented system was developed.

Visually Oriented System

During the first two phases of the current project, researchers developed and evaluated the "Visually Oriented System" (VOS), which was designed to be screen-reader-friendly for individuals who are blind but could also be operated by mouse and had a visual appearance very similar to regular ETS computer-based tests. For example, the screen had a "Next" button to advance to the next item. Text content was displayed with black Arial 10-point font on an off-white background. Items were from the content domains of reading comprehension, structure (grammar), writing, and listening comprehension—essentially, the set of items used in SVT2, plus one writing item. At the screen's bottom was information such as the time remaining and the item and section numbers ("1 of 3"). Various portions of the item display were implemented using HTML "frames," for example, one frame for the passage, one for the stem and choices, one for the "time remaining" indicator, one for the "testing tools" (i.e., "Next" button), plus other frames containing scripts to provide interactivity. Care had been taken during development of VOS to ensure that the frames containing scripts were the very last frames and that scripts within those frames were not voiced. The popular screen readers we used allow navigation both between and within such frames.

During these early two phases, the VOS was evaluated with 25 individuals representing a range of disability statuses (BL, LV, LD, DF, DB, ND).[20] The time that participants spent using the system varied widely, ranging from about 5 minutes to more than an hour. Researchers observed eight individuals, who had a basic familiarity with screen reader technology, as having great difficulty using the system. Moreover, the presence of the frames appeared to cause great difficulty for users who relied on screen reader technology. Indeed, users typically required frequent—sometimes constant—prompting from the researcher to know what to do next. VOS was much more usable by individuals who typically did not rely on screen reader technology. Individuals who were deaf-blind and relied completely on refreshable Braille (which was controlled by the screen reader) were almost entirely unable to navigate or enter responses. Reactions to the VOS interface by individuals who were sighted (i.e., deaf, nondisabled, or had learning disabilities) were generally positive.

It was uncertain how much of the difficulty for nonsighted individuals was due to factors such as lack of familiarity with the test structure and lack of user experience with screen reader (and refreshable Braille) technology or to problems in the design of the testing software. However, researchers believe that the use of HTML features such as frames had contributed significantly to the difficulty of access via screen reader. Navigating successfully between and within these frames seemed to require an unusually high level of screen reader (and possibly Braille display) expertise.

HTML-Form System

Based in large part on the difficulty of using VOS with a screen reader, the HTML-Form System (HFS) was developed to provide a screen-reader-friendly solution. It would use a Web browser (Internet Explorer version 5.5) and allow ready access, if necessary, via screen reader and Braille display. This new system simplified the structure of the test by “flattening” the test content by doing away with frames and linearizing the content. Specifically, this system consisted of a single Web page with labeled sections pertaining to the various test parts (sections, section directions, items, items directions, passage, question [stem], and choice list, etc.). The text content was displayed with black Times New Roman font of about 16 points on a white background. Part 1 contained one item in each of five sections (reading comprehension, structure, writing, quantitative comparison, and problem solving). The writing item used a text input box capable of receiving an extended essay; each of the other items in Part 1 provided a single-character text input box for test takers to enter the letter corresponding to their desired response to a single-selection multiple-choice test. Part 2 provided short, easy items intended to evaluate the usability of four different response options: dropdown box, single-character text input box, radio buttons, and check boxes.[21] At the page's bottom was a "Submit" button that, when clicked, would "submit the form" and save the data to a database.[22] Though potentially used with screen readers, HFS did not provide navigation instructions for a specific screen reader; it was simply a Web page that people were expected to navigate based on their own knowledge of Web content, Web browsers, and assistive technologies.

Development of Version 3 of the Self-Voicing Test

Guided in part by the findings of the evaluation of SVT2, the SVT system was enhanced and shortened to 5 items to become SVT3. Among the enhancements were the following:

1. The system was ported to a more stable software platform.
2. Auditory cues about progression through content were added.
3. Directions for items and sections were made reviewable and more extensive.
4. The system was modified to yield a more consistently reasonable result from the most basic command: "voice next paragraph" command (Control-Down Arrow).
5. The answer-confirmation sequence was modified to allow it to provide multiple-selection multiple-choice response formats.[23]
6. Greater use was made of a phonetic alphabet for single letters in commands ("Control-M as in 'Mike'").
7. Limited support for mouse was added.
8. The system was made available via an Internet connection, provided that the remote user had a recent version of Internet Explorer and the Microsoft Speech Software Development Kit version 4.0.[24]

SVT3 provided one item in each of 5 sections—reading comprehension, structure, listening comprehension, quantitative comparison, and problem solving.[25] Details of these and other enhancements are in Appendix A.

Part 3: Method for the Phase 3 Formative Evaluations

Overview

Fifteen adults, 2 to 4 from each of the six disability statuses—blindness (BL), low vision (LV), deafness (DF), deaf-blindness (DB), learning disability (LD), and no disability (ND)—participated in the final (Phase 3) formative evaluation of the systems. Each participant was administered from 2 to 15 items in each of one or two of the systems. Items came from the domains of reading comprehension, listening comprehension, structure (grammar), writing, and math. The participants provided feedback through interviews and focus groups. The key question was: "What are the strengths and weaknesses of the systems in meeting the access requirements of individuals with the six different disability statuses?"

Sample

The sample for the field test (Phase 3) consisted of 15 participants with BL (*n* = 4), LV (*n* = 2), LD (*n* = 3), DF (*n* = 2), DB (*n* = 2), or ND (*n* = 2).[26] Recruited participants were a convenience sample. All students were adults and were either college-bound (*n* = 2) or already had experience with higher education. Participants were given a number as an identifier (ID), which was established by the order in which the individuals participated in the study. Appendix D has greater detail on the sample.[27]

Computer System

The computer system for the project was a 700 MHz Pentium-based Compaq Armada M700 laptop computer running Windows 98. The system had 384 MB RAM, a 9 GB hard drive, and a 14-inch diagonal monitor with a 1024 x 768 pixel resolution and 16-bit color.[28] Audio was produced by two portable Yamaha speakers (model YST-M150). An Alva Satellite 544 refreshable Braille display, which provided a 44-character display panel, was available.[29]

Two popular screen readers—JFW (version 3.7) and Window-Eyes (version 4.0)—were available. During this Phase 3 study, the screen readers were used only with HFS.[30] The computer had connectors for modem and Ethernet.

Materials

During the third and final phase of the project, all three systems—SVT3, VOS, and HFS—were evaluated. VOS was the same system used for early phases of the project. HFS was new during this third phase.

Table 2 summarizes the three systems' features and their importance for different audiences.

Table 2

Summary of System Features Available During the Phase 3 Study

Feature	Estimated importance for individuals with disability status (high or medium)						Feature present in system (1 = yes; 0 = no)		
	BL	LV	LD	DF	DB	ND	SVT3	HFS	VOS
User features									
Base system provides text-to-speech	H	M	M				1	0	0
Supports optional use of screen reader	H	M			H		0	1	0[a]
Supports optional use of Braille display (under control of screen reader)	H	M			H		0	1	0[a]
Supports constructed responses	H	H	H	H	H	H	0	1	1
Supports keyboard-only navigation	H	H			H		1	1	1
Supports mouse-only navigation		M	H	H		H	0	1	1
Uses large font size	M[b]	H			M		1	1	0
Sample items ("Part 2") include various single-selection multiple-choice response formats	H	H	H	H	H	H	0	1	0
Content coverage									
Reading comprehension	H	H	H	H	H	H	1	1	1
Structure (grammar)	H	H	H	H	H	H	1	1	1

(Table continues)

Table 2 (continued)

Feature	Estimated importance for individuals with disability status (high or medium)						Feature present in system (1 = yes; 0 = no)		
	BL	LV	LD	DF	DB	ND	SVT3	HFS	VOS
Listening comprehension	H	H	H			H	1	0	1[c]
Mathematics	H	H	H	H	H	H	1	1	1
Writing	H	H	H	H	H	H	0	1	1

[a]As discussed earlier, the screen reader and Braille display capabilities of VOS seem to require high levels of technological expertise, probably too much to require of test takers, so these capabilities are rated as “not available” (“0”). [b]In this study, some individuals classified as blind rely somewhat on their sense of sight under some circumstances (in addition to relying heavily on speech output). [c]Listening Comprehension items were present in VOS but were not used in the Phase 3 study; both participants who used VOS in Phase 3 were deaf.

All delivery methods used disclosed items from the TOEFL test and from the GRE General Test quantitative section. Table 3 shows the kinds of items and response formats for the three systems. Table 4 shows response format alternatives that exist for HFS only.

Table 3

Number of Items Used for the Phase 3 Study in Each Content Area for the Three Systems

Main content	Testing system		
	SVT3	HFS	VOS
Reading comprehension	1	1	3
Structure	1	1	2
Writing	0	1	1
Listening comprehension	1	0	0[a]
Mathematics			
Quantitative comparison	1	1	3
Problem solving	1	1	3
Subtotal	5	5	12
Response formats for main content	Keyboard-only-based single-selection multiple-choice[31]	Single character entry for single-selection multiple-choice, plus text entry for writing	Keyboard- or mouse-based single-selection multiple-choice, plus text entry for writing

[a] The four listening comprehension items, though available, were not used for VOS; both participants who used VOS were deaf.

Table 4

Response Format Alternatives

Response format	Testing system (HFS only)
Drop-down box	1
Single-character entry	1
Radio buttons	1
Check boxes	1
Subtotal	4

Four single-selection multiple-choice items were common to all three methods. The reading item asked the test taker to identify the main idea of a passage. The structure item required the user to select from several options the best phrase to complete a sentence. The quantitative comparison item provided a graphic of a circle with labeled lines and points, and asked the test taker to indicate the relationship between the lengths of two lines. The item also provided a 114-word text description of the graphic. The problem-solving item required the test taker to solve an algebraic expression.

The listening comprehension item (which was common to SVT3 and VOS) provided a 9-second, 35-word conversation between a male speaker and a female speaker; in addition, a 7-second auditory description of the picture of the speakers was provided. The writing item common to HFS and VOS asked the test taker to write about a favorite academic experience.

Additional variety was encountered in the items found only in VOS. For example, one of the reading items provided a target sentence in the item stem and required the test taker to click a location within the passage to indicate where a target sentence should be added. Another reading item was a multiple-selection multiple-choice item that required two answers.

Additional Materials

Among the additional materials were (a) administrator screens that allow administrators (e.g., proctors) to launch different testing options, (b) a testing-selection tool that allows administrators to quickly identify which testing options might be most suitable for individuals with different characteristics, and (c) large-print and Braille materials.

Additionally, several data collection instruments were developed, notably the background questionnaire (administered before any use of any system), a post-method questionnaire (administered after the use of each system), and a post-session questionnaire. The background questionnaire, which was administered before the participants used any system, included questions about participants' educational and professional background, disability, and language experience and proficiency, as well as their prior experience with and reliance on alterations such as speech synthesis technology, human readers, and Braille technology. Appendix B has additional information on these materials.

Procedure for the Phase 3 Formative Evaluation

Overview of the Procedure

Each participant was administered from 2 to 15 items in each of one or two delivery methods (i.e., systems). Each participant was assigned to use one or two of the three delivery methods, all of which used Web browser technology: (1) SVT3 provided built-in text-to-speech capabilities and keyboard operation; (2) HFS used standard HTML form input elements (text input boxes, radio buttons, drop-down boxes, check boxes) and supported the optional use of text-to-speech (via screen reader software) and Braille (via refreshable Braille display); and (3) VOS, which was visually oriented (similar to current ETS computer-based tests) and was operable via mouse. Data also were collected through interviews and focus groups.

Assignment to a Treatment Set

Following the administration of the background questionnaire, examinees were assigned to one of three treatment sets, based on the examinee's background and experience and the researcher's understanding of which set of systems would be most desirable and suitable for the examinee. The goal of these assignments was to provide one or more systems that had a good possibility of being usable by the participant yet would allow, where feasible, comparisons between methods. The composition of the treatment sets was shaped by researchers' intentions to focus on evaluating SVT3 and HFS rather than VOS since much had already learned much about the strengths and weaknesses of VOS during earlier phases. When SVT3 was being administered, it was given first, even if it was not necessarily the optimal option (for example, in the case of participants who were nondisabled).

The three treatment sets were:

1. *SVT3 Treatment Set*. Administers SVT3 first, followed by HFS. This was the treatment type for all hearing participants (i.e., blind, low vision, learning disabled, nondisabled). This treatment set has three subtypes:

    a. HFS plain. This is for HFS users desiring neither screen reader nor Braille display.

    b. HFS with screen reader only. This was for individuals desiring screen reader but who were not users of refreshable Braille.

    c. HFS with screen reader and refreshable Braille. This was for users of both screen reader and refreshable Braille.

2. *VOS Treatment Set*. Administers VOS first, followed by HFS. This was the treatment type for all deaf (but not deaf-blind) participants. The listening comprehension items were not administered to these participants.

3. *HFS Treatment Set*. Administers HFS with screen reader and refreshable Braille display. This was the treatment type for participants who were deaf-blind. No detailed instructions specifically for screen readers or Braille displays were provided. Users had to rely almost entirely on their own expertise in these technologies.

If participants were using a screen reader with HFS, they were allowed to pick between JFW 3.7 and Window-Eyes 4.0.

Implementation of the Intended Procedure

The Core Materials Summary row in Table 5 shows the materials that were used with each of the participants.

Note that of the 15 participants, 11 individuals (all participants except those who were deaf or deaf-blind) used SVT3. Fourteen of the 15 used HFS. Two individuals, both deaf, used VOS.

Table 5

Background of Participants and Their Usage of Various Treatment Components

	Disability														
	BL				LV		LD			DF		DB		ND	
Identification number	20 [a]	28	31	35	32	40	27	29	39	37	38	33	34	30	36
Gender	M	M	F	M	M	M	M	F	M	M	F	M	F	F	M
Native English speaker	Yes	No	No	No	Yes	No	Yes	Yes	Yes	No	Yes	No	No	No	o
Age	40s	50s	20s	30s	50s	20s	20s	20s	40s	20s	20s	30s	20s	30s	30s
Education	4-C	2-C	2-C	Grad	4-C	Grad	4-C	4-C	4-C	2-C	HS	4-C	HS	Grad	4-C
Have you ever used a human reader?	Yes	No	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No
Core materials summary	SVT3 + HFSw/SR & RBD	SVT3	SVT3 + HFSw/SR & RBD	SVT3 + HFSw/SR & RBD	SVT3 + HFSw/SR & RBD	SVT3 + HFS	SVT3 + HFS	SVT3 + HFS	SVT3 + HFS	VOS + HFS	VOS + HFS	HFS + SR & RBD	HFS + SR & RBD	SVT3 + HFS	SVT3 + HFS
Used admin. screen	n/a	n/a	Voicing	Voicing	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Regular	Regular
Used testing-selection tool [b]	Yes	No	Yes	No	No	No	No	No	No	No	No	No	No	Yes	Yes
Used screen reader with testing-selection tool	Yes	n/a	Yes	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	No	No

Note. SR = Screen reader, RBD = Refreshable Braille display; 2-C = Two-year college graduate, 4-C = Four-year college graduate, Grad = Received graduate degree, HS = High school graduate. [a] The numbers do not start with number 1 because the sequence also encompasses earlier studies. [b] Actually called the "method-selection" tool during the study.

Inapplicable Data

In this study, when numbers of responses are reduced from what it appears they should be, data actually may be inapplicable rather than missing.[32] Some data were missing or excluded.[33] None of these problems is believed to compromise the tentative conclusions of the formative evaluation.

Limitations

The small and nonrepresentative sample limits generalizability of the findings regarding differences between disability type, language status, task type, and so forth. Among the limitations are: the small number of participants, the small number of disability statuses, a convenience (as opposed to representative) sampling scheme, and the small range of content types and response types. Therefore, the findings of this project should be interpreted with special care and considered as *exploratory* rather than confirmatory results.

Part 4: Results of the Phase 3 Formative Evaluation

This part describes the results of the Phase 3 formative evaluation. Major sections within this part pertain to: time reviewing the systems, opinions about specific features of the systems, opinions about use of the systems for important tests, comparisons between the SVT3 versus HFS systems, and opinions about the ideal testing system. Appendix B has additional findings.

Time Reviewing the Systems

The time that each individual took to review SVT3 ranged from 40 minutes to 152 minutes.[34] The time taken by the two reviewers (DF) of VOS were 16 minutes and 19 minutes. Also, the amount of time spent reviewing the HFS system ranged from approximately 5 minutes to 150 minutes.[35]

Opinions About Specific Features

The participants were asked several Likert-scale questions regarding the system's usability. Out of 15 individuals who participated in this study, 11 individuals reviewed SVT3, 14 individuals reviewed HFS, and 2 deaf individuals reviewed VOS. Appendix B has additional descriptions of participant opinions.

SVT3

The majority of the participants who reviewed SVT3 considered the system easy[36] with respect to understanding the synthesized speech (10 out of 11), the prerecorded speech for the listening comprehension items (10 out of 11), and the test directions from the system (10 out of 11). However, 1 blind individual (a nonnative speaker in his 50s) indicated that it was difficult[37] to understand the synthesized speech and the test directions from the system. Also, another individual (a nondisabled nonnative speaker in his 30s) chose “no opinion/undecided” to the question about the test directions provided by the researcher.

The users of SVT3 considered it easy to: go back and forth between the passage, question, choices, and navigation controls of the item (8 out of 11); confirm answers to items and then to proceed to the next item (11 out of 11); understand the descriptive speech for listening comprehension visuals (10 out of 10)[38]; and learn the key combinations for navigation and control (10 out of 11). However, 2 individuals with low vision considered it hard to go back and forth between the passage, question, and choices. Also, 1 individual (LD) considered it hard to learn the key combinations and other keyboard commands for navigation and control.

On the other hand, regarding the question about how easy it was to understand the text descriptions for math figures, among the 10 who answered this question, a total of 5 respondents (1 BL; 1 LD; 1 ND; 2 LV) found it difficult to understand the text descriptions for math figures. However, when the participants (9 individuals) were asked about the helpfulness of the text descriptions for math figures, 6 participants said “helpful,” while 2 individuals (1 BL and 1 LV) said “not at all helpful.”

Of those 5 individuals who responded, 4 considered the auditory descriptions (descriptive speech renderings) for listening comprehension visuals as “helpful,” while 1 individual with learning disability (a native speaker in his 20s) found it “not at all helpful.”

All 6 individuals (1 LV, 3 LD, 2 ND) who experienced the feature that highlighted in yellow each word as it was read by the synthesized speech, considered it very helpful.

Furthermore, a majority of the participants (10 of 11) considered the synthesized speech very or somewhat helpful, while 1 individual considered it “not at all helpful.” In fact, this individual (ND) expressed that the audio aspect of the system was a weakness or a problem.[39] Everyone considered it helpful to hear the word “more” after each paragraph and “end” at the end of passages, questions, and direction pages.[40,41]

HFS

Fourteen individuals reviewed HFS (3 BL, 2 LV, 3 LD, 2 DF, 2 DB, and 2 ND).

All 4 applicable individuals (3 BL, 1 LV) said that it was easy to understand the synthesized speech. The majority of users (12 out of 14) said it was easy to go back and forth between the passage, question, choices, and navigation controls of the item. One blind participant (a nonnative speaker) and 1 deaf-blind participant said that it was difficult to do so. Of the 10 applicable participants, 3 participants (1 BL, 1 DF, 1 DB) expressed that it was hard to type the essay into the text box (2 of these participants were using a screen reader with Braille display).

The majority (10 out of 12 applicable individuals) considered the radio buttons easy to use, except 2 blind participants, who said they were difficult to use. Regarding the question about how easy it was to use a drop-down box, 10 out of 12 applicable individuals considered it easy (the 2 deaf-blind individuals did not fully experience that response option), while 2 (1 DF, 1 BL) considered it hard. In particular, the blind individual considered it very hard to do so. Furthermore, all but 2 individuals said that it was easy to use the single-character input field. One deaf individual and 1 individual with low vision considered it hard. Also, when 12 applicable participants were asked to rate the item that used checkboxes to select more than one choice, 8 considered it easy, 2 (1 LV, 1 BL) considered it hard, 1 chose “no opinion/not decided,” and data for 1 was not available.[42]

VOS

The 2 participants—both deaf—who reviewed VOS said it was easy to understand the test directions provided by the researcher and to understand the system, and also considered it easy to type the writing essay into the text input box and to learn the key combinations and other keyboard commands for navigation and control. One individual considered it easy to go back and forth between the passage, question, and choices, and 1 individual chose “no opinion.” One considered the radio buttons easy to use, while the other one did not review the radio buttons.

Opinions About Use for Important Tests

Participants' Opinions Regarding SVT3

When the participants were asked whether they would recommend a testing system like SVT3 for very important tests such as for college or graduate school admissions, the majority (9 out of 11) would "recommend" it, while 2 nondisabled participants chose "would not recommend." Table 6 shows the participant responses based on their disability status.

Table 6

Participants' Responses About SVT3 Based on Their Disability Status

Question	Responses										
1. Disability status	Blind				Low vision		Learning disability			Nondisabled	
2. Identification number	20	28	31	35	32	40	27	29	39	30	36
3. Time reviewing SVT3 (in minutes)	9	72	84	58	75	57	40	85	120	81	152
4. Would you recommend a testing system like this for very important tests such as for college or graduate school admissions?	SR	SR	SR	R	R	R	R	SR	SR	WNR	WNR
5. I prefer this approach to a live human reader.	NAD	n/a	A	NAD	D	SA	n/a	n/a	n/a	n/a	n/a
6. I prefer this approach to the use of paper-and-pencil tests.	n/a	n/a	SA	n/a	A	A	n/a	NAD	SA	D	SD
7. I prefer this approach to hard-copy Braille tests.	A	n/a	D	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Note. SA = strongly agree, A = agree, NAD = neither agree nor disagree, D = disagree, and SD = strongly disagree; SR = strongly recommend, R = recommend, ND = undecided or no opinion, WNR = would not recommend.

When the 5 participants who had used a human reader were asked about their level of agreement with the statement "I prefer this approach to a live human reader," 2 participants

(1 BL, 1 LV) agreed or strongly agreed, while 1 (LV) disagreed. Two participants (2 BL) chose “neither agree nor disagree.”

Also, 4 out of 7 applicable participants (1 BL, 2 LV, 1 LD) preferred this approach to paper-and-pencil tests, 2 nondisabled individuals did not, and 1 person (LD) neither agreed nor disagreed. Of the 3 blind individuals who reviewed the system, 1 preferred this approach to hard-copy Braille tests, while the other participants did not.

The majority (8 out of 9 applicable individuals) expressed that they would recommend a system like SVT3 to individuals requiring speech output technology. One individual (LD) chose “no opinion/undecided” as her response. However, when the nonnative participants who were applicable (6 individuals) were asked whether they would recommend a system like this for English language learners, 2 blind individuals and 1 nondisabled individual highly recommended it, while 1 individual with low vision said “would not recommend” and 2 with learning disability stated “no opinion/ undecided.”

Table 7 shows what participants cited as the strengths of SVT3.

Table 7

Strengths of SVT3

Disability	Strengths
Blind	• The system allows users to complete a test independent of a human reader. • The system is easy to use and navigate, has a clear design, and is not verbose (in contrast to screen reader output). • The system allows easy navigation and one can interrupt the speech with another command. • The system allows easy navigation and it is easy to listen to.
Low vision	• The system allows a level playing field.
Learning disability	• The system’s commands were easy to use. • The system has options to repeat.
Nondisabled	• The system’s sections were clearly separated. • The system reads and highlights text. It would be great for English language learners because they are able to listen and relisten (avoids skipping content). • The system allows users to review instructions and change answers.

Participants' Opinions Regarding VOS

Both deaf participants who reviewed VOS recommended it for very important tests such as for college or graduate school admissions (see Table 8). Furthermore, participants agreed that they preferred this approach to using paper-and-pencil tests. Also, 1 nonnative participant recommended a system like this for English language learners (a question was asked only of nonnative participants).

Table 8

Participants' Responses Regarding VOS

Question	Responses	
1. Disability status	Deaf	
2. Identification number	37	38
3. Time reviewing VOS (in minutes)	16	19
4. Would you recommend a testing system like this (VOS) for very important tests such as for college or graduate school admissions?	Recommend	Recommend
5. I prefer this approach to the use of paper-and-pencil tests.	Agree	Strongly Agree

Also, the 2 deaf participants who reviewed VOS said they preferred the VOS system over the HFS system for taking tests. All information on one screen, less mouse work, and clear structure were mentioned as areas in which the VOS approach was superior to that of the HFS approach for testing.

Both participants either agreed or strongly agreed that they preferred the VOS approach to the paper-and-pencil tests. Also, they both strongly agreed that they preferred the VOS system over the HFS system for taking tests. Three interrelated areas of the VOS approach were mentioned as superior to that of the HFS approach for testing: The system contained all content on one screen, it involved less mouse work (such as scrolling), and it had a better visual layout (structure). On the other hand, 1 participant considered the font size of the VOS text to be too small and preferred the larger font size of HFS.

Individuals with deafness often have difficulties accessing written text even though their vision is unimpaired. The study did not specifically examine scores on items or have a sufficiently large sample of participants to detect such difficulties.

Participants' Opinion Regarding HFS

For HFS, 8 individuals recommended the system, while 6 (1 BL, 1 LD, 2 DF, 1 DB, 1 ND) would not recommend it for important tests such as for college or graduate school admissions. Table 9 shows the participants' responses based on the disability status. Individuals who were blind (2 out of 3), had low vision (2 out of 2), or had a learning disability (2 out of 3) recommended the system like HFS for important tests. The 2 deaf individuals chose "would not recommend."

Of the 5 participants who had used a human reader before, 3 preferred the HFS approach to using a live human reader. The 3 participants who preferred HFS over a human reader included the 2 low-vision participants and 1 blind participant. The other 2 individuals who did not prefer HFS over a human reader included 1 blind individual (a nonnative speaker in his 30s), as well as another blind individual who was undecided.

When users of paper-and-pencil tests were asked whether they preferred this approach over paper-and-pencil tests, 7 participants agreed, while 2 (1 ND, 1 DF) disagreed (see Table 9). When participants who could read Braille were asked whether they preferred this approach over hard-copy Braille tests, 3 (1 BL, 2 DF) agreed, while 1 (BL) disagreed.

Four of the 5 individuals who used speech output technology indicated that they would recommend a system like HFS to individuals requiring speech output technology, while 1 (ND) chose "no opinion/undecided."

HFS was recommended by all 4 users of refreshable Braille displays.

When the nonnative participants were asked whether they would recommend a system like HFS for English language learners, 5 chose "recommend," while 1 (a deaf nonnative speaker in his 20s) said "would not recommend."

Several comments on the strengths of HFS were mentioned (see Table 10).

Table 9

HFS Participants' Responses Based on Disability Type

Questions	Responses													
1. Disability type	Blind			Low vision		Learning disability			Deaf		Deaf-blind		Nondisabled	
2. Identification number	20	31	35	32	40	27	29	39	37	38	33	34	30	36
3. Time reviewing HFS (minutes)	86	150	16	21	20	15	29	9	5	6	71	95	73	23
4. Used screen reader	Yes	Yes	Yes	Yes	No	No	No	No	No	No	Yes	Yes	No	No
5. Used refreshable Braille display	Yes	Yes	Yes	No	No	No	No	No	No	No	Yes	Yes	No	No
6. Would you recommend a testing system like this for very important tests such as for college or graduate school admissions?	R	SR	WNR	SR	R	WNR	SR	SR	WNR	WNR	WNR	SR	R	WNR
7. I prefer this approach to a live human reader.	NAD	A	D	SA	A	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
8. I prefer this approach to the use of paper-and-pencil tests.	n/a	n/a	n/a	SA	A	SA	SA	SA	SA	D	n/a	n/a	A	D
9. I prefer this approach to hard-copy Braille tests.	SA	D	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	A	A	n/a	n/a

Note. SR = strongly recommend, R = recommend, ND = undecided or no opinion, WNR = would not recommend; SA = strongly agree, A = agree, NAD = neither agree nor disagree, D = disagree, SD = strongly disagree.

Table 10

Strengths of HFS

	Strengths
Blind	• The system allows independence. • The system contains a consistent structure. Liked the "single character input" response format, which is the response format most favorable for JAWS for Windows users.
Low vision	• Liked black text on white background (instead of white text on black background as in SVT3). • Part 1: The system was easy to navigate and contained a visible font size. • Part 2: It contains radio buttons, which were easy to use.
Learning disability	• Part 1: It was easy to read and to know how to answer (provided detailed direction). • Part 2: It contains radio buttons, which were easy to use. • Part 1: It was easy to navigate and had a large font size. • Part 2: "I like all [response] formats; [uses] multiple formats to keep it interesting."
Deaf	• Part 1: Detailed directions. • Part 2: Liked radio buttons. • Part 1: Allows users to use a mouse and scroll bar to review. • Part 2: Uses checkboxes and eliminates possibility for mistakes.
Deaf-blind	• This would help eliminate barriers to employment. • The system allows users to input answers without depending on a human scribe.
Nondisabled	• Part 1: The system was operated on a computer. • Part 2: It helps eliminate possibility for mistakes. • Part 1: The system allows user to scroll to review and to use a mouse. • Part 2: "Like radio buttons."

System Comparison: SVT3 Versus HFS

When the participants were asked to compare SVT3 and HFS, their opinions were mixed. First of all, 4 participants (2 BL, 2 LV) agreed that they preferred the SVT3 system over the HFS system for taking tests, while 4 (l BL, 1LD, 2 ND) disagreed (Table 11). Two participants (LD) chose "neither agree nor disagree."

Some of areas in which the HFS approach were cited as being superior to the SVT3 approach for testing were: no voicing (ND, LV, LD), mouse operation (ND), radio buttons (LD), easier to navigate (LD), shorter learning curve (BL), fewer keystrokes (LD), more personal control (LD), simplicity (LV), user friendliness (LV), clear speech (LV), and use of familiar screen reader (BL).

On the other hand, areas in which the SVT3 approach was cited as being superior to that of the HFS approach were: voicing (ND), highlighted word/sound tracking (LD), prerecorded audio (listening comprehension) (LD), synthesized speech (LD), description of visuals (LD), navigation convenience and speed (BL), fewer commands (LV), fewer crashes (than using JAWS for Windows with HFS) (LV), uniform interactive navigation (BL), and the lack of excessive verbosity (BL).

Of the 5 participants who had used a human reader before and who used both SVT3 and HFS: 2 individuals (1 BL, 1 LV) preferred both approaches over a human reader; 1 individual (BL) preferred HFS but not SVT3 over a human reader; 1 individual (BL) did not prefer SVT3 over a human reader but had no opinion or was undecided about HFS; and, finally, 1 individual (BL) did not prefer either SVT3 or HFS over a human reader.[43]

Table 11

Preference for SVT3 Over HFS

Question	Responses														
Disability	Blind				Low vision		Learning disability			Deaf		Deaf-blind		Nondisabled	
Identification number	20	28	31	35	32	40	27	29	39	37	38	33	34	30	36
Gender	M	M	F	M	M	M	M	F	M	M	F	M	F	F	M
Language	N	NN	NN	NN	N	NN	N	N	N	NN	N	NN	NN	N	NN
I prefer the SVT3 system over the VOS for taking tests.	SA	(Did not use HFS)	D	SA	SA	A	D	NAD	NAD	n/a	n/a	n/a	n/a	D	D

Note. N = native speaker of English, NN = nonnative speaker of English, SA = strongly agree, A = agree, NAD = neither agree nor disagree, D = disagree, SD = strongly disagree.

The Ideal Testing System

All the participants were asked to discuss what the major features of an ideal testing system would be. The following were mentioned: clear speech, quick and easy navigation from questions to answers, selection review, ability to independently navigate text, access to Braille, option to turn sound on/off, usable with mouse, option to change font size, hard-copy Braille, refreshable Braille with screen reader, less text, hotkeys like the shortcuts of SVT3, an option to change answers, voicing with browser, and use of ZoomText.[44]

Other Findings

Appendix B has details on additional findings.

Part 5: Highlighted Topics

This part—in a Question-and-Answer format—highlights topics that bear upon efficient and effective creation and delivery of accessible tests. Most of the issues are directly informed by this project.

Participant Recommendation of a Solution

Question: Did all participants find a solution that they could recommend?

Answer: Notwithstanding a variety of usability problems, almost all participants (13 of 15) would recommend at least one of the delivery methods for high stakes tests, such as those for college or graduate admissions. See Table 12.

HFS was arguably the system with the broadest (though not always deepest) appeal, having been recommended by persons in 5 of the 6 different disability statuses.

Thus, the three delivery methods appeared to address many of the requirements that participants expected in computer-based delivery methods for very important tests, such as for college or graduate school admissions. Six of the 15 participants would recommend both delivery methods that they used, and an additional 7 participants would recommend one of the methods that they used. For operational use, however, a number of usability issues must be addressed. Furthermore, there are test creation and security issues to address.

Table 12

Did Participants Recommend a Delivery System for "Very Important" Tests?

ID	Disability	SVT3	HFS	VOS
20	BL	Yes	Yes	n/a
28	BL	Yes	n/a	n/a
31	BL	Yes	Yes	n/a
35	BL	Yes	No	n/a
32	LV	Yes	Yes	n/a
40	LV	Yes	Yes	n/a
27	LD	Yes	No	n/a
29	LD	Yes	Yes	n/a
39	LD	Yes	Yes	n/a
37	DF	n/a	No	Yes
38	DF	n/a	No	Yes
33	DB	n/a	No	n/a
34	DB	n/a	Yes	n/a
30	ND	No	Yes	n/a
36	ND	No	No	n/a

Only 2 individuals, 1 deaf-blind individual and 1 nondisabled individual, did *not* recommend any delivery method that they used. The deaf-blind individual used only one delivery method (HFS, with screen reader and Braille display), and the nondisabled individual used two methods (SVT3 and HFS). As noted, deaf-blind individuals had great difficulty accessing the test content. Had the nondisabled individual been administered VOS, he might well have recommended that method because of its mouse operation and visually oriented interface.

Among the 9 individuals who used both SVT3 and HFS, 4 of the 5 individuals with blindness or low vision preferred SVT3, while 3 of the 5 nondisabled or learning disabled individuals preferred HFS.

The 2 deaf-blind participants, who had great difficulty accessing the content, used HFS with a refreshable Braille display under the control of screen reader software (JAWS for Windows). One deaf-blind individual recommended HFS for important tests; the other did not.

Key Features

Question: What are the key elements or features of solutions?

Answer: The study has led the authors to conclude that the three delivery systems, in aggregate, addressed many of the most important *elements* needed to develop tests for individuals with diverse disability statuses—blind, low vision, learning disabled, deaf, deaf-blind, and nondisabled. Among these elements are:

1. Synthesized speech via built-in capabilities (SVT3) or screen readers (HFS)
2. Refreshable Braille (HFS)
3. Intuitive visual design (VOS and somewhat for the other systems)
4. Alternative content (text descriptions of figures and math expressions [SVT3, HFS, VOS], auditory descriptions of visuals in listening comprehension stimuli [SVT3, VOS])
5. Constructed response entry (HFS, VOS)
6. Keyboard operation (SVT3 and nearly complete in others)
7. Mouse operation (VOS, HFS, partial in SVT3)
8. Large base font sizes (SVT3, HFS)

Each delivery system included a sufficient number of critical elements that more than half of its users would recommend it for "very important" tests. For example, both of the deaf individuals who used VOS recommended it, 9 of the 11 people who used SVT3 recommended it, and 8 of 14 people who used HFS recommended it. In all, 13 of 15 participants used a system that they could recommend.

Each delivery system had special strengths. SVT3 showed the strengths of its text-to-speech interface with people with visual disabilities. VOS appeared quite effective for individuals who were deaf, and would likely prove effective for individuals who are nondisabled.[45] HFS was the only delivery system that supported refreshable Braille and screen reader use and, again, had the broadest (though not always deepest) appeal.

Even with the strengths of these systems, there were many usability problems identified. Participants suggest fixing these problems before using the systems in operational high-stakes testing situations. Appendix B has details on some of these perceived weaknesses.

Reuse in Expediting Item and Test Development

Question: What kinds of *reuse* are most feasible in developing tests that are accessible to individuals with disabilities?

Answer: In software development, if not test development, reuse is an important strategy for ensuring efficiency. Ideally, for example, one would be able to take the text content for a reading comprehension item and reuse it in several presentation formats—visually displayed text for test takers who are deaf or nondisabled, synthesized speech for test takers who are blind, and Braille for individuals who are deaf-blind or blind. The ideal situation would be to have nothing peculiar in a specific presentation format that would require a different text source for each of the presentation formats. The idea of a high level of reuse is implied in the ideal of "universal design," in which something is designed to be usable by virtually everyone.

Reuse of Templates and Content in SVT3

This project made extensive and successful reuse of both item *templates* and *content* across packaging variants of SVT3.[46] This meant, for example, that an edit to an item content file or an item directions file for a reading comprehension item became immediately available to any of the SVT3 variants on the system that used that reading comprehension item.[47]

Reuse of Content Across Presentation Formats and Audiences

Another kind of reuse that was only partly achieved in this project was reuse of content across presentation formats and audiences. To understand some of the different manifestations of this kind of reuse, let us consider an ideal situation in which the same test item content can be used (i.e., reused) regardless of whether it is displayed visually for deaf or nondisabled test takers, in synthesized speech for test takers who are blind, or in Braille for test takers who are deaf-blind or blind.

This kind of reuse is perhaps most feasible in content areas that rely on text alone and that do not rely on special symbols, such as math symbols. Theoretically, extremely high levels of reuse would be feasible in such a topic area. Yet, when it comes down to developing items for these three different test formats, other task variables come into play and make it less practical to reuse the item content.

Reuse in an "Insert Text" Reading Comprehension Item

For the project, the VOS system was developed with the intent to make an "insert text" TOEFL reading comprehension item reusable for two different audiences: sighted users (e.g., deaf, learning disabled, nondisabled) who generally use a mouse, and nonsighted users of screen reader technology who use a keyboard. This particular "insert text" item required the test taker to determine where to insert a target sentence (presented in the stem) within a paragraph in order for the revised paragraph to make the best sense. With a little practice, this item format is easy for sighted people to use. However, the format was virtually unusable for nonsighted screen reader users, at least under our evaluation conditions. Perhaps with considerable practice a very competent user of screen reader technology might learn to use the system. However, test designers might wonder if performance on such an item might be unduly influenced by construct-irrelevant factors such as screen reader expertise.

In implementing the SVT3 "Basic" system,[48] this "insert text" item was adapted in several ways. It was adapted to become a straightforward single-selection multiple-choice item. The five choices described the locations in which the target sentence might be placed: "After the first sentence"; "After the second sentence"; and so on to the fifth sentence. Another adaptation of the item removed the passage's last three paragraphs, since the target locations were only within the first paragraph. The adaptation was seen as reasonable, since a sighted VOS user would instantly see the possible insert locations were all in the first paragraph and nonsighted users of SVT3 Basic might unnecessarily waste time feeling that they had to listen to all four paragraphs of the passage simply because they were there to be read.[49] In summary, a test item that worked well for sighted users of VOS needed to be adapted for use in environments that may be used by nonsighted individuals (e.g., SVT3 Basic and HFS Basic), thereby making it hard to reuse the same item content for all of this project's audiences.

Summary on Reuse

Perhaps it is best to think of an "item" as a family of item variants, each of which is intended to be suitable to a set of audiences and delivery formats. The variants within a family typically will share a common topic as well as much of their verbiage. Ideally, one variant will share (reuse) media resources (text, graphics) in common. This approach is consistent with current developments in automatic generation of items based on test purpose, analysis of the construct, population characteristics, cognitive models, and so forth (Bejar, 2002; Irvine &

Kyllonen, 2002) as well as an ECD design approach, which represents assessment designs and their parts as reusable objects (Mislevy, Steinberg, & Almond, 2003). Nevertheless, it is also important not to rely on overly optimistic estimates of reusability of test content. As noted, the researchers observed that many format-specific adaptations to test content needed to be made for different delivery systems and audiences. In planning for larger-scale implementations of accessible computer-based tests, it is important to carefully examine assumptions of very high amounts of test content reuse across different presentation formats.

Extensible Markup Language (XML) technologies provide a way of representing test content to promote such sharing and reuse. XML encourages the separation of content and format, which promotes reuse by facilitating the rendering of the same content in multiple formats. XML provides opportunities to develop rich, well-structured representations of the "default" test content (consisting of text, audio, video, images, etc.), of the alternative content (descriptive text for audio, video, or images), and of other information needed to render data in many different formats. (See Appendix H for an example that shows how XML markup could be used to separate and label chunks of content that a testing application could display or not display, depending on the access requirements of the user.)

High priority needs to be given to making adaptations that will promote fairness and validity, which may, at times, reduce the amount of reuse that is possible. Given the limitations on reuse that are likely to be encountered on certain fronts, it is important to look well beyond simple reuse of content; for example, reuse of specifications, personnel, plans, and procedures can also be extremely valuable.

Screen Readers and Braille Displays

Question: How effective and capable are key assistive technologies, particularly screen reader and Braille display technologies?

Answer: A few observations seem relevant here.

1. The quality of these assistive technologies has increased greatly in the past several years.
2. There is increased attention to the importance of ensuring the presence of alternative content, such as text descriptions of graphics (Chisholm, Vanderheiden, & Jacobs, 1999).
3. There is also recognition of the need for improvements to content viewers, notably Web browsers, to ensure good access through assistive technologies. For example, the *W3C*

User Agent Accessibility Guidelines (Jacobs, Gunderson, & Hansen, 2002) provide guidelines to developers of Web browsers for making them work well with assistive technologies.

4. Evaluation of the effectiveness of these technologies is challenging, especially in the case of Braille displays. For example, when observing a deaf-blind or blind individual interacting with HFS with a screen reader, it was common to find the visually displayed text to be dozens of lines of text out of synch with the synthesized speech.[50] A non-Braille-reading researcher had no way to immediately determine how much information and of what kind was being displayed on the refreshable Braille display. In such an environment it was difficult to distinguish between problems and "features" and, if there was a problem, to determine whether it had to do with the nature of the authored content, the Web browser, the screen reader, Braille display, the user's level of technological expertise, or some interaction between factors. Future research needs to take these challenges into account.

5. Some improvements in screen reader performance might be obtained either by developing configuration files that are suited to a testing application and/or by allowing test takers to configure screen readers for themselves. Our participants who used screen reader technology did not request or suggest modifying the configuration files. We surmise that use of these files is considered an advanced feature with which our sample had limited experience.

Challenges in Deaf-Blind Access

Question: What were the causes of difficulty in deaf-blind individuals' access to HFS?

Answer: Some of the possible difficulties include:

1. *Inadequate time to become familiar with the structure of the test document*. Some difficulties might have been due to unfamiliarity with the structure of the test, including the location and order of passages, questions (stems), choice lists, form controls (single-character input field, radio buttons, etc.), use of internal links, and the spatial relationship between labels and input fields, among others.

2. *Less-than-optimal familiarity with the various navigation modes of the screen reader (JAWS for Windows)*. Many screen reader users are able to function well with familiar content relying on one or two modes, and may not be familiar with modes needed to access other content or applications (C. Earl, personal communication, October 31, 2001).

3. *Large document size*. Some of the navigation problems could have been due to the large size of the document, which included the whole test, including all items and the test, section, and item directions. A screen reader may take many seconds or up to a minute or more to both load and do preprocessing on the document. The screen reader may permit the user to navigate during the loading and preprocessing period, but there is a greater likelihood of navigation anomalies during that period. Furthermore, such loading and preprocessing periods may be experienced several times for a given page, depending on how the user changes navigation modes for that page (C. Earl, personal communication, October 31, 2001). For this and other reasons, the size of individual HFS documents should be reduced. (Note that since some JAWS for Windows users *were* successful in operating HFS, this suggests that large document size was not the cause of all the navigation problems.)

4. *Low speed of person-to-person communication*. A general challenge in examining deaf-blind access to HFS concerned the limited speed of person-to-person communication. During one interview with a participant who was deaf-blind, the researcher communicated via spoken word to a hearing sign language interpreter, who then signed the message to a deaf tactile interpreter. The deaf tactile interpreter then conveyed the message to the deaf-blind participant via tactile (hand-to-hand) signing. The deaf-blind participant then signed his response in the reverse direction, except that the deaf-blind participant could sign in the air to the hearing signer, who would then voice the message to the researcher.[51] The time necessary to communicate by this process reduced the number of communications that we could accomplish in a set amount of time.

Tutorials and Directions

Question: What special challenges are presented for the development of format- and audience-specific tutorials and directions?

Answer: There is a critical need to gain experience—through research—in actually producing effective tutorials and directions that focus on integrating the basics of a particular computer-based test and the accessibility features that have been introduced.

Tutorials and directions would be tailored to both the delivery format and the intended audience. For example, SVT directions for a nonsighted person might emphasize the use of keystrokes for entering responses, while directions for sighted users might emphasize mouse operation.

The strategy for directions was different between SVT3 and HFS. Because HFS could, in principle, be used several different ways—visually, auditorily (via synthesized speech), and via refreshable Braille—directions were quite generic. Users were expected to be knowledgeable in the use of the Web browser and, if applicable, the screen reader and the Braille display. (More than one person suggested providing directions on what screen reader commands to use.) On the other hand, SVT3 directions were optimized for keyboard operation and auditory interface for test takers with visual disabilities, particularly blindness. Directions for SVT3 were labor-intensive to create, and focused largely on a single category of disability (visual disabilities). If the Self-Voicing Test had the option of turning voicing on and off (for sighted users who want to rely on it occasionally), then directions more suited to visual operation should be considered.

It would seem feasible to modularize pieces of directions and eventually develop systems that have tutorials and directions that are tailored to the person using the system and to the delivery method.

Improving Speech Quality

Question: How can the speech output of the Self-Voicing Test be improved?

Answer: Speech quality is a key issue for the Self-Voicing Test. Some participants seemed genuinely distracted by instances of poor pronunciation. Even a single word mispronounced was sometimes cited by at least one participant as being a significant distraction; for example, when the synthesized speech pronounced the word “CON-tent” as “con-TENT.” Others complained that all the speech sounded somewhat muffled. Although our sample of nonnative speakers generally had a relatively high level of English proficiency, one individual who rated himself as having a middle level of English proficiency had an exceptionally difficult time understanding the synthesized speech. This is not surprising, but it is something to keep in mind when planning speech capabilities that are to be used by individuals with limited English proficiency.

A variety of approaches need to be examined to improve the quality of speech. Sensible next steps would be to try different voices and intonation settings within the same engine, or replace the speech engine being used. Other kinds of improvements might be implemented through program logic, Audio Cascading Style Sheets, MathML, and better quality control to identify trouble spots. Parts of a solution might include greater use of phonetic alphabets (e.g., "Control-Shift-D as in Delta" or "Control-Shift-C as in Choice") or spelling out words known to be troublesome (e.g., "… unfamiliar, spelled u-n-f-a-m-i-l-i-a-r, customs," which otherwise might be perceived as "familiar") in the body of the text. Making such modifications to content that is intended for synthesized speech output may limit easy reuse of the same content for a different presentation format (e.g., visually displayed text), but other kinds of reuse (such as of people, plans, specifications, etc.) might still allow efficient development of accessible tests. New or forthcoming speech technologies may be helpful (AT&T Labs Natural Voices, n.d.). The use of prerecorded human speech will probably continue to be essential for some applications (e.g., listening stimulus in listening comprehension items) but may also be worthy of consideration more broadly in some settings.

The Role of Screen Reader Technology

Question: Why are screen readers an important part of the solution for disability access to computer-based tests, and what are likely requirements for their successful use?

Answer: At the time of the Phase 3 study, the screen readers provided two major capabilities not available in the Self-Voicing Test version 3 (SVT3): access to constructed response formats and refreshable Braille support. Such capabilities are likely to be necessary for access to computer-based tests.

For screen readers to be useful in high-stakes testing, the user must have a rather high level of screen reader expertise and must have plenty of time to become familiar with the test format. Ideally, test takers should be provided with reference materials (Braille or electronic format, as appropriate) to help users navigate and control the testing application.

In addition, the content must be authored to take into account the diverse ways in which a screen reader might be used. For example, if punctuation is critical to understanding content, there may need to be warnings about the necessity of having punctuation voicing turned "on," or there may need to be some other way to signal the presence of punctuation with words or special characters.

In implementing a screen-reader-based solution, priority should be given to working with experts to develop training materials that would help users interact with testing materials from popular screen readers (e.g., JAWS for Windows or Window-Eyes). Such training might even include some configuration of the screen reader to optimize it for the testing application. The configuration file might be saved for reuse in later sessions.

HTML and Browser Technologies

Question: Is HTML and Web browser the best combination for delivering screen-reader-friendly tests?

Answer: Notwithstanding the many advantages in delivering tests as HTML files that can be viewed with Web browsers, other combinations also may have important potential.

One participant who was deaf-blind indicated that HTML code viewed with a Web browser was not a good way to develop tests. Two deaf-blind individuals independently indicated that a much more screen-reader-and-Braille-friendly method would be to deliver the test in a Microsoft Word document. The test taker would simply place an agreed-upon marker (such as triple asterisks) beside their selected option for multiple-choice items and type in the text for constructed response items. An obvious disadvantage of this approach is that one essentially loses the automatic scoring that is such a major advantage of many computer-based tests. An advantage of test delivery via an MS Word document is that screen reader vendors have expended considerable effort in making their tools work well with word processing software.

Delivery environments other than HTML/Web browser might be considered. A consultant who is blind emphasized the improved screen reader access that might result from developing in a regular programming environment, such as C, C++, or Visual Basic. Delivering a test as an HTML document in a Web browser may not give the test developer a high level of control over the test taker's experience. For example, the test developer is prohibited from programming the system to accept certain keystrokes (e.g., Down Arrow or Control-P) for test delivery functions since these are reserved and intercepted by the Web browser for its own purposes. This is a major reason for the Self-Voicing Test having triple-key combinations, such as Control-Shift-P instead of Control-P to go to the "passage." In summary, if screen-reader-friendly test delivery is required, it may make sense to look at development environments other than basic HTML Web browser delivery, although future browsers may overcome some of these difficulties as well as having new accessibility features (Jacobs, Gunderson, & Hansen, 2002).

Braille and Large-Print Figure Supplements

Question: Will Braille and large-print figure supplements continue to be important?

Answer: The researchers' observations suggest that Braille and large-print figure supplements will continue to be useful—if not essential—in many cases. The descriptions of math visuals, while helpful to some people (with and without visual disabilities), do not appear to provide a fully adequate replacement for the figure supplements. This issue needs more investigation. One individual (BL) made a strong appeal for great care in the development of clear text descriptions of math visuals. A strategy needs to be developed that would provide a variety of supplements, such as text descriptions of figures and math, large-print and raised line Braille figures, and human readers where they are necessary. Other innovative resources for access to figures and math also might be investigated, such as the use of audio tones to represent graphical plots (Gardner, 1999; IMS Global Learning Consortium, 2002; Rothberg & Wlodkowski, 2003) and tactile graphics with touch-activated audio cues (Landau, 2002).

Verbosity and Time Limits

Question: Given that the participants in the study were not under much time pressure, what new issues are likely to arise in the use of a system like SVT3 as time limits (albeit extended time) are imposed to replicate operational testing situations?

Response: Among the impacts of limited time on use of a system like SVT3, we expect an increased user interest in controlling the *verbosity* of messages and directions, and possibly higher severity estimates for usability problems in general. Verbosity refers to the wordiness of what is presented by the system to the user.

Screen readers such as JAWS for Windows and Window-Eyes provide the user with considerable control over verbosity of message. For example, JAWS for Windows 3.7 allows the user to determine whether messages are long or short and whether 18 other attributes (such as control group name, control type, control name, dialog name, dialog text, error messages, status information, Smart Help messages, Tool Tip, position information) are spoken. The system provides default settings for beginner, intermediate, and advanced levels, but allows the message length and the 18 attributes to be configured in any manner. Additional verbosity control is available for HTML pages. Among the setting parameters for these are: skip past repeated text on new pages, say link type, text link verbosity, graphics verbosity, image map link verbosity, graphics link verbosity, new frame indication, identify “same page” links, and identify tables. The system also allows user control over speaking of punctuation and formatting. Configuration settings can be saved in a configuration file for later use.

With time limits reimposed, users of a system like SVT3 also would likely want greater control over certain aspects of verbosity. For example, most pages are introduced with the following phrase: “Press Control-Down Arrow or press Control-M as in Mike for a list of navigation options.” Even though the user can interrupt the voicing by invoking a navigation command, some participants wanted to have the system quit saying that. While the inability to have the system not present that message was not seen as serious, the addition of time pressure might change that. Other content-specific directions also could seem excessive. For example, each of the section directions contained instructions about the structure of the items (e.g., whether they contained a passage or not). Item instructions typically have some redundancy with the section directions and provide specific instructions about navigation. For example, the “Item Directions for Reading Comprehension Item 1” says: “Item 1 refers to a reading passage. At any time, you can navigate to the passage (using Control-Shift-P as in ‘Passage’), the item (using Control-Shift-Q as in ‘Question’), or choice list (using Control-Shift-C as in ‘Choice’).” This kind of reminder about keystrokes is appropriate while people are learning or if keystrokes are different than those used earlier, but may interfere with examinee success in other circumstances. An important priority for future research would be to identify what kinds of control the test taker needs to have over key aspects of their testing experience, such as the verbosity of the auditory rendering of test content.

An Opportunity for Easier Navigation in SVT

Question: How might navigation be made easier in SVT?

Answer: Features to provide *easier navigation* between major components of an item might make a big difference in usability. Generally, people liked the Self-Voicing Test’s use of “hot-key” commands such as Control-Shift-P to go to the passage, Control-Shift-Q to go to the question (stem), and Control-Shift-C to go to the choice list. When they were concentrating on test content, however, they sometimes forgot those commands. They could readily go to the command menu for a reminder, but even this interruption could break their concentration. A new version of the Self-Voicing Test might modify the navigation scheme to ensure that a single simple command—Control-Down Arrow—allows the test taker to navigate through an entire item. In SVT3, Control-Down Arrow speaks the next paragraph but does not navigate across component boundaries. Under a revised navigation scheme, the boundary between the passage and the question, for example, would be "soft" so that Control-Down Arrow could be used to continue from the end of the passage through the boundary to the first paragraph of the question. Thus, under such an approach, Control-Down Arrow could allow one to navigate forward and backward through an item without hitting component walls that required hard-to-remember commands to continue.

A Role for HFS

Question: What is a reasonable role for HFS or approaches like it?

Answer: Advantages of HFS include its simplicity and its basic accessibility through assistive technologies such as screen readers and Braille displays. Whomever is able to effectively use these technologies with Web pages can, in principle, enjoy features such as Braille support and constructed response formats, which are still absent from SVT3.[52]

It is interesting to note that although resources devoted to HFS were only a small proportion of what were devoted to SVT3 (or even VOS), there was still a range of individuals who would recommend it. Specifically, of the 9 individuals who used both HFS and SVT3, 4 preferred HFS over SVT3, compared to 3 who preferred SVT3 over HFS and 2 who were undecided. HFS-like tests might be delivered over the Web as practice tests or, with adequate safeguards, for other high-stakes purposes.

HFS could be a good platform for developing prototypes for item creation and test assembly capabilities. For example, one can imagine automating the generation of HFS response format variants—one variant for sighted test takers who prefer radio buttons for answer choices and another variant for blind or deaf-blind test takers who prefer single-character input boxes (because this may be more reliable when used with a screen reader).

However, lest we think that approaches like SVT are unnecessary, one should note that SVT3 was preferred by 4 of the 5 individuals with blindness or low vision who used both SVT3 and HFS.

Part 6: Discussion and Conclusions

This final part provides discussion and conclusions. It begins with a discussion of three large-scale imperatives that, if acted upon, would promote effective and efficient progress toward more accessible computer-based testing.

Three Imperatives

The following imperatives respond to trends in validity, inclusion, and technology that were mentioned in this report's introduction. These imperatives integrate knowledge and developments that have come about in the past few years, as well as understandings from earlier portions of the study.

The three imperatives are:

1. Take steps to provide both accessibility and validity.
2. Commit to making consistent progress in implementing inclusive systems.
3. Use new technologies to leverage organizational resources.

Take Steps to Provide Both Accessibility and Validity

The design of computer-based tests needs to ensure that accessibility features do not undermine validity. Sometimes it is fairly obvious when an accessibility feature conflicts with validity. For example, to provide a spell checker on a test of spelling ability would tend to invalidate the results of the test. On the other hand, some features, such as the use of the readaloud accommodation on tests of reading, are quite controversial. Specifically, for example, the readaloud accommodation is prohibited on the reading assessment of the National Assessment of Educational Progress (NAEP), yet the accommodation is permitted in some state assessments of reading achievement (U.S. Department of Education, 2003, p. 71).

ECD is a promising approach for helping ensure the *validity* of inferences arising from accessible computer-based (and other) tests (Hansen, Forer, & Lee, 2001; Heath & Hansen, 2002; Mislevy, Steinberg, & Almond, 2003; Hansen, Mislevy, & Steinberg, 2003). ECD frames each assessment or test as embodying an evidentiary argument. That argument gains its substantive meaning by its relationship to real-world knowledge as well as the values that support the field of educational measurement—such as validity and fairness. Within such a framework, a major goal is to *remove*, to the extent possible, any *unfair disadvantages* resulting from the examinee's disability, typically through altering features of the test performance situation, while at the same time preventing *unfair advantages* for the person receiving the altered test. The argument for an assessment might be summarized as including: (a) a claim about a person possessing (or not possessing) a certain proficiency, (b) the data (e.g., scores) that would result if the person possessed (or did not possess) the proficiency, (c) the warrant (or rationale, based on theory and experience) that tells why the person's possession (or lack of possession) of the proficiency would lead to occurrence of the data, and (d) "alternative explanations" for the person's high or low scores (i.e., explanations other than the person's possessing or not possessing the proficiency).[53] The existence of alternative explanations that are both significant and credible might indicate that validity has been compromised. An example of an alternative explanation for low scores by an individual with a disability would be that the individual is not able to receive the test content because there is a mismatch between the test format (e.g., visually displayed text) and the individual's disability (e.g., blindness). An example of an alternative explanation for high scores would be that the accommodation eliminates or significantly reduces demand for some aspect of the targeted proficiency. Such an alternative explanation might be highly credible if an individual with a spelling disability were offered the use of a spell checker as an accommodation on a test intended to measure spelling proficiency. Recognizing such alternative explanations and determining how to address them can sometimes be difficult.[54]

Hansen, Mislevy, and Steinberg (2003) have focused on building argument structures that might help anticipate and address these alternative explanations, particularly as they relate to test takers with disabilities. Using a Bayes net tool, they have developed runnable models of these arguments, based on very simple components. While such models cannot automatically make key decisions, they can illuminate the nature of the decisions and help assessment designers think through the sometimes competing goals of assessment designs.

Addressing such alternative explanations may require making changes or alterations to the *administration* activities, such as those that ensure appropriate kinds of uniformity of testing conditions and the appropriateness of the evidence gathered and accumulated; the *pre-administration* activities, such as those that determine who is eligible for certain alterations in testing conditions; and *post-administration* activities, such as reporting of test results. ECD provides an assessment designer with a framework or schema for making the argument explicit.

An ECD approach can allow one to better anticipate how a change in one aspect of the argument will affect the other parts of an argument, and what actions need to be taken to promote the coherence of the whole argument. By guiding the development of more robust representations of the evidentiary argument associated with testing individuals with disabilities, the ECD approach might more quickly highlight areas of concern in any prospective changes to tests or formats.

The challenge of developing a coherent argument for tests given under altered conditions for individuals with disabilities is similar to that of developing such an argument for *language adaptation* (e.g., adapting a test originally created in English to French). In language adaptation, the goal is to arrive at an "equivalent" version of some test in a different language. Since not all components of a test may be directly translatable, adaptation of a test into another language may involve rewriting or replacing certain items (Hambleton, 1994; Stansfield, 1996).

Any framework for assessment design should help ensure that tests measure what they were intended to measure, rather than other (construct-irrelevant) influences. Care is needed to ensure that the scores resulting from the use of any operational testing system are valid for the intended purposes (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Elliott & McKevitt, 2000; Hansen, Forer, & Lee, 2001; Heath & Hansen, 2002; Willingham & Cole, 1997; Willingham et al., 1988).

Commit to Making Consistent Progress in Implementing Inclusive Systems

Technology and great design concepts are not enough to ensure progress toward accessible computer-based tests. Also needed is organizational commitment. It may be useful to think of the process of improving the inclusiveness and accessibility of computer-based tests for people with disabilities as occurring in four levels that represent stages of inclusiveness (in the sense of universal design) and of integration of system functionalities.

- Level 1: Conventional Approaches
- Level 2: Transition Approaches
- Level 3: Special-Purpose Systems
- Level 4: Inclusive Systems

It is important to note that one does not need to entirely leave one level to progress into the next. A testing organization may have its testing operations distributed across all four levels, and elements of Level 1 are likely to appear in higher-level operations.

Level 1: Conventional approaches. Solutions at the first level consist of the conventional or current ways of administering computer-based tests and their existing alternatives or alterations. Specifically, they include conventional computer-based tests (with or without alterations) plus additional relevant formats (e.g., regular and large-print paper, audiocassette, Braille) and figure supplements (e.g., Braille and large print) that become relevant when the current computer-based testing system is inappropriate for a test taker with a disability. Some components that are currently found at this level, such as the current generation of CBT and audiocassette, might eventually be phased out, while others (human reader, tactile graphic supplements, large-print paper) may continue to exist in some degree long into the future.

Level 2: Transition approaches. Solutions at the second level consist of a varied and loosely integrated collection of systems or components that lay the foundation for more integrated approaches (Levels 2 and 3). Many of these components may consist of commercial off-the-shelf products. This level includes approaches like HFS or the use of test content in a Microsoft Word file. Such approaches might have the advantage of providing quick development time and ready use of screen reader and refreshable Braille display. But they might lack automated scoring capabilities and would therefore rely on strong institutional support for rapid human scoring and reporting. Approaches that are part of this level would include capabilities that are new, are undergoing exploration, or offer interim solutions pending longer-term solutions at the higher levels.[55] For example, using test content in a Microsoft Word document that can be accessed via screen reader and Braille displays at the present might be considered such an interim approach since it offers important capabilities (e.g., constructed response formats, Braille output) not currently available in a more conventional computer-based testing format. Another example of such an approach would include an SVT3-like system if it were used operationally. Some of the approaches at Level 2 might consist of incremental but important improvements beyond the Level 1 approach; for example, instead of using audiocassette to present prerecorded audio, one might use a Digital Talking Book player to present prerecorded audio.[56] Such a change might greatly improve the navigability of the content. For a test taker who is blind, such a change might be coupled with test taker entry of answers into a word processor file (using a screen reader) or with use of a human scribe. While such incremental improvements might fall short of the systems described for Levels 3 and 4, they could still constitute significant improvements in service to test takers with disabilities.

Level 3: Special-purpose systems. Solutions at the third level include systems intended almost exclusively for specific groups of individuals with disabilities. Examples of important accessibility features for individuals with visual and learning disabilities might include: keyboard navigation, modifiable font sizes, screen enlargement, color modification, speech synthesis, highlighting of spoken text, refreshable Braille support, descriptions of nontext content, audio descriptions for video, mouse support, support for diverse selected and constructed responses, and settable testing times. Such systems might typically include relatively small portions of a testing organization's item pool. An enhanced version of the Self-Voicing Test might constitute an example of such an approach.

Level 4: Inclusive systems. Solutions at the fourth level would involve the functionalities of one or more systems with capabilities similar to those of Level 3, but would closely integrate those functionalities with systems that serve individuals *without* disabilities. Achieving this level of integration would involve designing tests and items for accessibility from the beginning, with due consideration of the validity implications of potential alterations. Such systems can be improved over time by integrating approaches that have been tried out at Levels 2 and 3. The great diversity of disability and other factors will probably dictate that a large testing organization never has all its operations at this fourth level, yet it does seem possible to serve a significant majority of test takers with disabilities through such systems.

For the purposes of this report, the key concept is that a strategy for CBT accessibility should allow for multiple approaches that are suited to different access requirements. Some approaches are evaluated, become operational on a small scale, and may become part the system that serves the vast majority of test takers. Other approaches may remain small because they are built for a very small number of individuals, perhaps just one individual. Even solutions developed for a single individual contribute to the base of knowledge about what works under various conditions. Finally, though the four levels are based primarily on *test delivery* features, important improvements in efficiency might be obtained at any level by greater up-front attention to accessibility during *assessment design* and during *item and test development*.

Use New Technologies to Leverage Organizational Resources

New technologies create opportunities for new kinds of testing (Bennett, 1999b). Wise use of technology can bring to bear the dramatic increases in CPU speed and decreases in the cost of computing power to serve all test takers, including individuals with disabilities. The ability to share resources across vast distances on high-speed networks can increase levels of reuse of resources, thereby lowering the cost of assessment delivery (including updates and adaptations). "Smart" assessments could adapt to the access needs of students with disabilities, taking account of the validity implications of potential adaptations. On the other hand, unwise use of new technologies can raise even greater access barriers; for example, the pervasive use of video and audio presentations, if not accompanied by appropriate alternative representations, can present barriers for individuals with visual and/or hearing disabilities.[57] Thus, new technologies can increase both opportunities and pitfalls in the quest for accessible computer-based tests.

Following are some of the most promising technologies.

XML (Extensible Markup Language) and related technologies are some of the most important and promising technologies for promoting the accessibility of computer-based tests and other learning products and services. XML encourages separation of content from formatting, thereby increasing the potential for representing content once and automatically rendering it in any of several formats. For example, text content might be rendered as Braille, visually displayed text, or synthesized speech.

Some important technological innovations are applications of XML. For example, Synchronized Multimedia Integration Language, or SMIL (pronounced "smile"), which is an application of XML, allows synchronization of virtually any kind of media that can be presented by a computer (audio, video, text, graphics, etc.). An interesting application of SMIL is the Digital Accessible Information SYstem (DAISY) standard for Digital Talking Book (DTB) technology. A DAISY application allows a user to navigate quickly between chapters, paragraphs, or other chunks of content designated by the content developer. Auditory navigation cues and keyboard or keypad access facilitate use by individuals who are blind or have other print-related disabilities. The "talking" aspect of a DTB is typically implemented via prerecorded audio files that can be built into the talking book, although some DTB players are also *compatible with screen readers,* such as JAWS and Window-Eyes. Screen reader access to DTBs may also facilitate output of content as refreshable Braille. Some new DTB players provide *built-*

in speech synthesis or allow one to mute the speech synthesis and use a screen reader if one desires. DTB players generally come as either stand-alone hardware (with an integrated keypad) or software-only (for which screen reader compatibility is applicable). DTB technology, thus, can provide an interesting mix of capabilities that encompass *both* direct access *and* access via compatibility with assistive technology. DTB players include features useful for sighted and partially sighted individuals, including mouse access, modifiable text color and size, and visual highlighting of text content while it is being voiced. DTB was designed to present book content, but it might be adapted to present test content, perhaps treating an individual item as a "chapter." While the DTB is essentially a format for *presenting* information, a DTB player could be integrated into software applications that include facilities for *capturing and processing* student response data. Additionally, DAISY version 3.0 and beyond may include additional modules for handling features such as math, dictionaries, video, workbooks, and testing. The DAISY Consortium would need to clarify the process for endorsing proposed modules in order to extend DAISY in this way. (Personal communication, George Kerscher, March 24, 2004).[58] See the DAISY Consortium Web site for information about DTB technology (www.daisy.org). Another application of XML is the Accessible Testing System, which was developed by gh, LLC, and pilot tested by ETS (gh, 2003).

Other kinds of open-source or proprietary formats based on XML could be used to store a rich representation of items such that they could be rendered in many diverse ways that are suited to different audiences with or without disabilities.

An important area for investigation concerns special formats for making technical content, such as math and chemistry, more accessible (IMS Global Learning Consortium, 2002). Many of the techniques being explored now are not widely available, but they may become important, especially in assessments of higher-level math, music, or other subjects that employ complex or special notations.

Other technologies that are of potential importance for developing accessible computer-based tests include standards and guidelines, such as those for instructional and test-related information, such as the IMS Global Learning Consortium (IMS Global Learning Consortium, 2002; Heath & Hansen, 2002), for browsers and other software for rendering Web content (Jacobs, Gunderson, & Hansen, 2002), or for information technologies more generally (Architectural and Transportation Barriers Compliance Board, 2000).

It should be noted that computer-based testing systems should allow a significant amount of configurability, flexibility, and control (IMS Global Learning Consortium, 2002; Jacobs, Gunderson, & Hansen, 2002). This is true not only because it may be desirable to give the test taker considerable control over his or her own testing experience, but also because validity considerations may require that the test developer, administrator, or proctor be able to turn certain features off or on to better ensure the validity of the inferences arising from using the test (Heath & Hansen, 2002).

Conclusions

It is feasible to build at least prototypes of computer-based testing systems that meet a great many accessibility requirements of many people with disabilities. Continued research and development is needed to refine, integrate, and augment the capabilities needed for operational computer-based testing systems. The small number of participants and other factors limit the strong conclusions that can be drawn; nevertheless, following are some tentative conclusions and hypotheses:

- Individuals with visual disabilities and particularly individuals with deaf-blindness face numerous challenges in accessing computer-based test content, and considerable attention should be given to assisting them in overcoming access barriers.
- It appears that test content for speech or Braille rendering of test content for high-stakes testing may continue to require adaptation (perhaps including dropping certain items).
- Refreshable Braille is a critical supplement for many individuals who read Braille, but hard-copy Braille needs to be maintained as an option, unless it is clear that it is not needed.
- Tactile graphics are not likely to be entirely replaced by text or audio descriptions of graphics.
- Individuals with disabilities other than blindness, such as many individuals with learning disabilities and deafness, also encounter access challenges, but may not face basic usability challenges as severe as individuals with blindness or deaf-blindness.
- Careful provision needs to be made to ensure that students with disabilities have adequate materials and time for becoming familiar with the formats being used.

Recommendations

Following are recommendations for testing organizations that wish to develop computer-based testing systems that are more accessible:

1. Carry out efforts to improve the accessibility of computer-based tests within a general framework that recognizes validity as a preeminent technical consideration. ECD shows promise as being able to serve such a role.

2. Become familiar with emerging standards and guidelines for accessibility of computer-based materials and, as necessary, adapt or extend them for your purposes. Existing standards and guidelines can help identify high-priority features as well as point to effective strategies for their implementation.

3. Investigate new technologies that will promote the efficient delivery of accessible test content. XML and related technology can provide rich representations of test content that can be rendered in diverse formats with greater ease. Improvements in other technologies such as Web browsers, media players, assistive technologies (e.g., screen readers, Braille displays), and tools for making graphical or technical content accessible deserve continued attention.

4. Involve people with disabilities in the design and formative evaluation of accessible systems, and conduct research on operational use of such systems. Such research should yield information about: (a) the usability of accessibility features (including those associated with practice and familiarization materials), (b) how to most efficiently provide new features (e.g., whether to build in certain accessibility features or to rely on assistive technologies), and (c) the validity of providing alterations in particular settings and populations.

5. Develop strategies for reuse that include reuse of content but also encompass schemas, models, or other higher-level representations of items, as well as reuse of personnel who are skilled in disability access issues.

References

Abraham, R. G., Plakans, B., Carley, M., & Koehler, K. (1988). *Screening/training prospective nonnative teaching assistants: The Iowa State University experience*. Unpublished manuscript.

American Council of the Blind. (2000). *American Council of the Blind Resolution 2000-45*. Retrieved October 20, 2004, from http://www.acb.org/resolutions/res2000.html - 2000-45

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

AT&T Labs Natural Voices. (n.d.). *Demos [of AT&T Natural Voices Text-to-Speech Engine]*. Retrieved October 20, 2004, from http://www.naturalvoices.att.com/demos/index.html

Architectural and Transportation Barriers Compliance Board. (2000). *Electronic and information technology accessibility standards*. Retrieved October 20, 2004, at http://www.access-board.gov/sec508/508standards.htm

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds), *Item generation for test development* (pp. 199-217). Mahwah, NJ: Lawrence Erlbaum.

Bennett, R. E. (1999a). Computer-based testing for examinees with disabilities: On the road to generalized accommodation. In S. J. Messick (Ed), *Assessment in higher education: Issues of access, quality, student development, and public policy* (pp. 181-191).

Bennett, R. E. (1999b). Using new technology to improve assessment. *Educational Measurement: Issues and Practice*, *18*(2), 5-12.

Chisholm, W., Vanderheiden, G., & Jacobs, I. (Eds.). (1999). *Web content accessibility guidelines*. Retrieved May 5, 1999, from the World Wide Web Consortium site: http://www.w3.org/TR/WAI-WEBCONTENT

Coyne, K. P., & Nielsen, J. (2001). *Beyond ALT Text: Making the Web easy to use for users with disabilities.* Fremont, CA: Nielsen Norman Group.

Elliott, S. N., & McKevitt, B. C. (2000). *Testing accommodations decisions: Legal and technical issues challenging educators or "good" test scores are hard to come by*. Paper presented at the AERA, New Orleans. Retrieved October 20, 2004, from http://www.wcer.wisc.edu/testacc/publications/AERApaper4-2000.htm

ETS. (1998). POWERPREP software: Preparation for the computer-based TOEFL test (Version 1.0) [Computer software]. Princeton, NJ: Author.

ETS. (1999). Test of English as a foreign language sampler [CD-ROM]. Princeton, NJ: Author.

ETS. (2000). *TOEFL 2000-2001 Information bulletin for supplemental TOEFL administrations.* Princeton, NJ: Author.

ETS. (2002). *Information about testing accommodations.* Retrieved October 20, 2004, from http://www.ets.org/disability/info.html

Gardner, J. (1999). *TRIANGLE: A mathematics scratch pad for the blind*. Retrieved October 20, 2004, from http://dots.physics.orst.edu/triangle.html

gh. (2003). *gh, LLC and Educational Testing Service working to provide historic computer-voiced standardized test for people with visual disabilities.* Retrieved October 20, 2004, from http://www.ghbraille.com/microsoftnewsrelease.html

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, *10*(3), 229-244.

Hansen, E. G., Forer, D. C., & Lee, M. J. (2000). *Interim report for the Self-voicing Test project*. Princeton, NJ: ETS.

Hansen, E. G., Forer, D. C., & Lee, M. J. (2001, April 13). *Technology in educational testing for people with disabilities*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Seattle, WA.

Hansen, E. G., Lee, M. J., & Forer, D. C. (2002). A "self-voicing" test for individuals with visual disabilities (Research Note). *Journal of Visual Impairment & Blindness*, *96,* 273-275.

Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2003, April). Evidence-centered assessment design and individuals with disabilities [Abstract]. In E. G. Hansen (Organizer), *Assessment design and diverse learners: Evidentiary issues in disability, language, and non-uniform testing conditions.* Symposium conducted at the annual meeting of the National Council on Measurement in Education, Chicago, IL. Retrieved on October 20, 2004, from the ETS Web site: http://www.ets.org/research/conferences/aera2003.html#evidence

Heath, A., & Hansen, E. (2002). Guidelines for testing and assessment. In IMS Global Learning Consortium (Ed), *IMS guidelines for developing accessible learning applications*.

Retrieved October 20, 2004, from http://imsproject.org/accessibility/accv1p0/imsacc_guidev1p0.html#1312344

IMS Global Learning Consortium. (Ed). (2002). *IMS guidelines for developing accessible learning applications*. Retrieved October 20, 2004, from http://imsproject.org/accessibility/accv1p0/imsacc_guidev1p0.html

Irvine, S. H., & Kyllonen, P. C. (Eds). (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

Jacobs, I., Gunderson, J., & Hansen, E. (Eds.). (2002). *User agent accessibility guidelines: W3C Recommendation 17 December 2002*. Retrieved October 20, 2004, from the World Wide Web Consortium site: http://www.w3.org/TR/UAAG10

Landau, S. (2002). *Development of an audio/tactile accommodation for standardized math tests for use by individuals who are blind or visually impaired* (Final Report for Small Business Innovative Research [SBIR] Phase 1 Project ED-01-PO-3667.) Brooklyn, NY: Touch Graphics.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds), *Item generation for test development* (pp. 97-128). Mahwah, NJ: Lawrence Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. Mack (Eds.), *Usability inspection methods* (pp. 25-62). New York: John Wiley & Sons.

Ostrow, M. (n.d.). *Screen reading: JAWS for Windows*. Austin, TX: University of Austin, Texas. Retrieved October 20, 2004, from the Library Services for People With Disabilities Web site: http://www.lib.utexas.edu/services/assistive/jaws.html

Phillips, S. E. (2002). Legal issues affecting special populations in large-scale testing programs. In G. Tindal & T. M. Haladyna (Eds), *Large-scale assessment programs for all students* (pp. 109-148). Mahwah, NJ: Lawrence Erlbaum.

Rothberg, M., & Wlodkowski, T. (2003). *Making educational software and Web sites accessible: Design guidelines including math and science solutions*. Retrieved October 20, 2004, from the WGBH Educational Foundation Web site: http://ncam.wgbh.org/cdrom/guideline/

Stansfield, C. W. (1996). Content assessment in the native language. *Practical Assessment, Research & Evaluation*, *5*(9). Retrieved October 20, 2004, from http://pareonline.net/getvn.asp?v=5&n=9

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Thurlow, M. L., Quenemoen, R. F., & Lehr, C. A. (2002). *Access to computer-based testing for students with disabilities* (Synthesis Report 45). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Lutkus, A. D., & Mazzeo, J. (2003). *Including special-needs students in the NAEP 1998 Reading assessment. Part I: Comparison of overall results with and without accommodations*(NCES 2003-467).Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (1988). *Testing handicapped people*. Needham Heights, MA: Allyn and Bacon.

Wilson, K. M., & Lindsey, R. (1999). *Validity of global self-ratings of ESL speaking proficiency based on an FSI/ILR-referenced scale* (ETS RR-99-13). Princeton, NJ: ETS.

Notes

[1] Screen reader programs are commercial software packages that allow a person who is blind (or has certain other specific disabilities) to interact with a variety of other software—such as word processors, spreadsheets, and Web browsers—using synthesized speech output.

[2] A refreshable Braille display has hardware that raises and lowers Braille dot patterns on command from a computer. Users can read the Braille dot patterns by moving their fingertips across the Braille panel, much as they would read the Braille dot patterns on paper.

[3] Except where noted otherwise, this document uses the term *alteration* as a general term for changes from default (or standard) testing conditions provided to a student with a disability. It includes changes that alter the proficiency being measured (sometimes called modifications) and those that do not (sometimes called accommodations). Phillips notes that the term *accommodation* unfortunately has also come to refer to "any assistance given to a student with a disability during the administration of an assessment" (Phillips, 2002, p. 123); this document attempts to avoid that usage.

[4] Raised line Braille diagrams use Braille to label the relevant features of the diagram just as a regular (print) math diagram has text labels for key features.

[5] Some of the disadvantages of human readers include inconsistent quality of reading, test-taker anxiety and embarrassment at having the reader re-read material, reader mistakes in recording answers, fatigue caused by the slowness and intensity of this reader/test-taker interaction, a greater need for extra testing time, and so forth. In one study, an individual who is blind elaborated on one source of anxiety associated with a human reader: "When you tell the reader that the answer to a multiple-choice question is 'b' and he asks 'b?' it makes you wonder. Is he asking me because he didn't hear me correctly or does he know that 'b' is *not* the correct answer and he is trying to give me a hint? It makes you anxious to wonder about things like that. It's much better if a person can take the test independently."

[6] More generally, assistive technologies consist of software or hardware that has been specifically designed to assist people with disabilities in carrying out daily activities, e.g., wheelchairs, reading machines, devices for grasping, text telephones, vibrating pagers, etc. (Jacobs, Gunderson, & Hansen, 2002).

[7] Voice recognition is not among the offered assistive technologies. Individuals who might use voice recognition software may be eligible to use a scribe (e.g., someone to write the answers).

[8] Improvements in text-to-speech technologies are yielding increasingly natural-sounding synthesized speech output.

[9] "To [use JAWS for Windows to] navigate in the Windows environment and in specific applications one must use the keyboard and one of the cursor modes. There are five cursor modes:

"1. PC Cursor mode: This is the default cursor mode for JAWS. The cursor stays in the main part of the application you are using. If you need to perform a function that is not part of the application, you will need to use a hot key to access that function.

"2. JAWS Cursor mode: This cursor acts like a mouse pointer does. You can access any part of the screen without having to use hotkeys. Press the minus key on the number keypad to activate it.

"3. Virtual PC Cursor: This cursor loads by default in HTML documents. The Virtual PC cursor works in much the same way that the PC cursor does, using JAWS navigation commands to read pages containing links and frames. One difference is that, unlike the PC cursor, there is no visual indication on the area of the screen that is being affected. Another difference is that text is read in one uninterrupted stream, whether the text appears on the screen or not.

"4. Braille Cursor: For use with the ALVA Braille Terminal. See the instructions 'ALVA Braille Terminal—The Basics' for more information.

"5. Invisible Cursor: In Invisible Cursor mode only speech will follow the cursor movements." (Ostrow, n.d.)

[10] Even well-designed content, such as Web content that adheres to the W3C Web Content Accessibility Guidelines (Chisolm, Vanderheiden, & Jacobs, 1999), can still present accessibility problems to some users.

[11] Later, this report presents a hybrid approach. However, the contrast between these two methods of speech-output access to electronic content—"screen reader" and "self-voicing application"—is sufficient for our current purposes.

[12] The Heath and Hansen quote (2002) uses the term accommodation in its common meaning of referring to any assistance given to a student with a disability during the administration of an assessment (Phillips, 2002, p. 123).

[13] Content was from the Law School Admissions Council.

[14] The system only used source materials that, in their original format, employed a regular single-selection multiple-choice response format (i.e., picking one choice from a short list of choices), as opposed to multiple-selection multiple-choice or other response formats.

[15] By contrast, all ETS tests are timed, even if extra testing time is allowed as an accommodation.

[16] The SVT2 reading comprehension, structure, and listening comprehension parts consisted of disclosed (nonsecure) items from the TOEFL PowerPrep® test preparation software (ETS, 1998), except for one structure item, which was drawn from the *TOEFL 2000-2001 Information Bulletin for Supplemental TOEFL Administrations* (ETS, 2000). Items in the math part consisted of disclosed items from the *GRE Big Book* (ETS, 1996).

[17] The system was a Compaq Armada M700 with 384 MB RAM running Windows 98 and Internet Explorer 5.5.

[18] Additionally, it could be presented via refreshable Braille, if supported.

[19] Many of these improvements have resulted from use of MSAA, as mentioned earlier.

[20] The sample included 25 participants with these disability statuses: BL (*n* = 6), LV (*n* = 7), LD (*n* = 7), DF (*n* = 1), DB (*n* = 2), or ND (*n* = 1). Recruited participants were a convenience sample. All students were adults and were either college-bound or already had experience with higher education. Of the 25 participants, 8 were nonnative speakers of English. Thirteen were female, and 11 were male.

[21] These items used researcher-developed content that was intended to be quite easy from a content standpoint. For example, a “quantitative comparison item” might require the test taker to determine the relationship between two quantities: “25” versus “100 divided by 4.” The correct choice to this item would be an assertion that they are “equal.” Except for the check-box item (which was a multiple-selection multiple-choice item), all items in Part 2 were single-selection multiple-choice items.

[22] As indicated in the materials preparation section, there were minor variants of the HFS.

[23] The capability for multiple-selection multiple-choice items was not used during the Phase 3 study.

[24] As noted below, the Internet option was not practical for listening comprehension items.

[25] SVT was packaged to meet several possible purposes within the project. The "full" SVT3 package (15 items) was split into two parts: SVT3 Listening Comprehension, which included listening comprehension (4 items), and SVT3 Basic, which included everything else. The exclusion of listening comprehension in SVT3 Basic made it possible to administer the test over the Internet; the multimedia files of the listening comprehension items made their use over the Internet impractical. SVT3 Basic was further split into SVT3 Reading Comprehension and Structure and SVT3 Math, and—most importantly—SVT3 Short, which provided one item in five sections (reading, structure, listening, quantitative comparison, and problem solving). Generally throughout this document (particularly outside this materials preparation section), the term *SVT3* refers to *SVT3 Short*.

[26] Nondisabled participants were included in the sample as they would help establish a baseline or point of comparison for usability as well as help validate the successful application of principles of universal design that are intended to make a system usable by virtually everyone, not just individuals who have disabilities. Also, the inclusion of nondisabled individuals would help evaluate the extent to which features designed to increase accessibility for people with disabilities might create usability barriers for people without disabilities.

[27] It was challenging to recruit participants for the various phases of this project. It was a special challenge to identify nonnative users of English who had the appropriate disability status and educational background.

[28] On a few occasions the HFS or VOS software was run over the Internet, which generally appeared to give virtually an identical experience as running with all software entirely on a laptop. The use of software over the Internet was ceased after some experience with page-loading delays due to an overloaded local network.

[29] For the individuals who were deaf-blind, most of the study was conducted with the participants' Alva Satellite 570 refreshable Braille display, which appeared to function the same as the 44-character display of the Alva 544 except for having a 70-character display instead of a 44-character display.

[30] ZoomText Level 1 screen magnification software also was available and could have been used with any of the three systems, but was not used during the field evaluation.

[31] Mouse operation was actually available but was not used during the Phase 3 study.

[32] For example, researchers asked only questions for which participants had a basis for providing an informed answer. If a participant did not use a certain system feature, then the participant was not asked how helpful it was (although that participant could receive a different question, such as how helpful a certain feature would have been).

This notion of applicability is particularly important regarding HFS, since it had features that were seen only by segments of all HFS users. The term *applicable*—usually used when referring to participants—is used to refer to individuals who experienced the feature. For example, individuals who were either deaf or deaf-blind were not asked about how easy the synthesized speech was to understand—although the speech was emitted—since they would not be able to experience it because they could not hear it. In this case, those individuals with certain disabilities were considered not applicable to the question. It can be assumed that those who reviewed a particular feature were considered applicable to answer the question about that feature.

[33] Following are missing and excluded data:

Of the 11 individuals who used SVT3, 7 viewed (and generally answered) items in all five sections. The other 4 participants (#20, 28, 30, 31) viewed items in fewer than all five sections.

One blind participant (#28) reviewed SVT3 but not HFS.

One blind participant (#20) reviewed SVT2 during an earlier phase of the project and reviewed SVT3 for only 9 minutes. He was instructed to respond to the posttreatment questionnaire with regard to the SVT3; he was not asked questions that could not be answered on the basis of his experience with SVT3 alone.

Of the 14 individuals who used HFS, 8 individuals viewed (and generally answered) items in all five sections of Part 1 and all items in Part 2 (alternative response formats). Others (#20, 27, 28, 31, 32, 33, 34) viewed items in less than five sections. The first exposure to HFS by 4 of the individuals who used HFS Basic (see the Material Preparation section) was followed by later exposure to HFS Single Character Short (i.e., the standard variant for this Phase 3 evaluation). One person (#29) reviewed only HFS Basic. The differences between the variants of HFS were believed to be insignificant for the purposes of this study.

All data from one individual (BL) was excluded entirely from the study because we were unable to obtain a posttreatment interview.

With the time and resources provided, neither of the deaf-blind participants (#33, 34) was able to respond to more than two items.

[34] These figures include time spent discussing usability problems but do not include initial time spent receiving instruction about the system (this was typically less than a minute per system). Additional time discussing usability problems sometimes occurred after the timed period, such as during an interview. One participant's record was excluded from the range given here because that blind individual took only 9 minutes to review SVT3; he had, however, spent a longer period reviewing SVT2 in an earlier phase of the project. The 9 minutes included several minutes reviewing changes from the previous version. The participant was only asked about those system features that he had experienced in the 9 minutes. Also, note that there was an individual (#36, ND) who took 152 minutes to review the system.

[35] Note that 2 deaf individuals reviewed HFS in approximately 5 minutes.

[36] The term *easy* in this context includes both *very easy* and *easy* in a scale that also includes *hard*, *very hard*, and *no opinion/undecided*.

[37] The term *difficult* in this context includes both *very hard* and *hard* in a scale that also includes *easy*, *very easy*, and *no opinion/undecided*.

[38] TOEFL listening comprehension audio stimuli are accompanied by still pictures that give information about the setting of the conversation. Auditory descriptions of the pictures were added to the audio track to make the information in the pictures accessible to individuals who were unable to receive or process the visual stimuli.

[39] Furthermore, all participants (10 out of 11) except 1 with low vision who said "no opinion/undecided," considered that having the option of using natural-sounding *prerecorded* human speech in addition to synthesized speech *would be* helpful (7 considered it very helpful, while 3 [2 BL and 1 ND] considered it "somewhat helpful"). The 1 individual, who said "no opinion/undecided," commented that he chose that rating because he felt it would take too long to finish the test if one kept using the prerecorded speech in addition to the synthesized speech; the very presence of prerecorded human speech would cause it to be used frequently, thereby wasting precious time.

[40] These features were implemented in SVT3 to address the problem that participants often failed to realize that there was *more* to hear in a component, particularly a passage.

[41] Other possible features were discussed during the session. For example, 6 individuals (2 LV, 3 LD, and 1 ND) considered that *large-print* (hard-copy) math figures would have been "helpful." In addition, 1 blind individual thought that hard-copy *Braille* math figures would have been "very helpful." (Large-print and Braille math figures were offered, where relevant, but only after the participant had a chance to try to answer the question without them.) All applicable participants except 1 individual with low vision considered the font size of the visually displayed text appropriate.

[42] The difficulty with the checkbox format may have been due in part to insufficient clarity about the fact that the user was allowed to select *more than one* option.

[43] Options to this question were "yes," "no," and "no opinion/undecided."

[44] ZoomText Level 1 is screen magnification software. Although it was available on the system, none of the participants in this Phase 3 study used it. The person who cited this is a person whose vision had been deteriorating. He had placed greater reliance on tools like ZoomText in the past and by the time of our study placed greater reliance on speech output tools such as screen readers.

[45] Although this study did not administer VOS to nondisabled individuals, VOS has a visually oriented interface similar to regular ETS computer-based tests, which have proved quite successful with many users, most of whom are nondisabled.

[46] These variants include, for example, SVT3 (i.e., SVT3 Short), SVT3 Basic, SVT3 Listening Comprehension, SVT3 Math, and SVT3 Reading Comprehension and Structure.

[47] This kind of reuse is largely a feature of the structure of the C3 test delivery platform that was used for SVT3. Use of the C3 platform made the development and maintenance of variants much more feasible than it was with the non-C3 programming structure used for SVT2 and SVT1.

[48] SVT3 Basic was a variant that used this particular item.

[49] The same basic structure used for this item in SVT3 Basic was also used in HFS Basic. The insert text item was not used either in SVT3 (i.e., SVT3 Short) or HFS (i.e., HFS Short).

[50] Evidently, there are ways to bring them back into synch, but to ask the user to do this repeatedly would have been very disruptive.

[51] In some situations the tactile signer was hearing, which reduced by one the number of steps in each direction.

[52] In this respect it is the only current approach that has some viability for individuals who are deaf-blind and rely on refreshable Braille displays.

[53] The appropriateness of this simple characterization is based on assumptions such as the stability of the purpose, the reference population, and the applicable set of related societal values and world knowledge. A more detailed characterization might give explicit focus to comparisons between the arguments (or facets of the argument) for two or more groups or populations (e.g., individuals with blindness versus individuals without any disability). It might also recognize proficiency and performance as having more than two levels (good, poor).

[54] Allowing a student to have test content read aloud to them (via human reader, prerecorded audio, or synthesized speech) is an important alteration of which the appropriateness is recognized as being difficult to judge (Phillips, 2002, pp. 127-128).

[55] For example, it may be more feasible to implement accessible versions of linear tests (that present items in a fixed sequence) than accessible versions of computer-adaptive tests (that present items in a sequence that varies depending upon the test taker’s response to earlier

items). In such a case, the accessible linear version might be considered an interim approach that would be taken until the computer-adaptive version becomes practical.

[56] Players generally come as either stand-alone hardware (with an integrated keypad) or software-only, which generally relies on the computer's keyboard for input.

[57] For example, video clips should be accompanied by captions, audio descriptions, and a collated text transcript (Chisolm, Vanderheiden, & Jacobs, 1999; Jacobs, Gunderson, & Hansen, 2002).

[58] As of the time of writing (March 2004), George Kerscher serves as Secretary General of the DAISY Consortium.

[59] By contrast, all ETS tests are timed, even if extra testing time is allowed as an accommodation.

[60] The C3 system makes use of XML data formats for storing test content and other important test information.

[61] This is consistent with the practice of TOEFL tests but inconsistent with the practice in the GRE General Test.

[62] We use the term *phonetic alphabet* to mean a system to spell things out over voice communications. Such systems, also called radio/spelling/telephone alphabets, use words instead of letters (e.g., Alpha, Bravo, ...Yankee, Zulu).

[63] Nondisabled participants were included in the sample to help establish a baseline or point of comparison for usability. The inclusion of nondisabled individuals was also intended to ensure the focus on principles of universal design that might help make the system usable by virtually everyone, not just individuals who have disabilities. Finally, the inclusion of nondisabled individuals would help evaluate the extent to which features designed to increase accessibility for people with disabilities might create usability barriers for people without disabilities.

[64] All nonnatives had a non-English first language, but several had learned English very early and most had been in English-speaking regions for more than 10 years.

[65] The participants who were deaf or deaf-blind were not asked, because speaking ability was not considered a good index of their English proficiency. The interview question was from Abraham, Plakans, Carley, and Koehler (1988), cited in Wilson and Lindsey (1999).

[66] The student identifier (ID) was established by the order in which the individuals participated in the study. The numbers do not start with number 1 because the sequence also encompasses earlier studies.

[67] In the cases of Regular Print (option 12) and Large Print (option 11), the Voicing Administrator Screen prompts the user to contact the researcher, who would then print the version from MS Word. A Large-Print version of the test was actually available on the system but not used.

[68] Two of the options, which were not available on the regular administrator screen, are Regular Print and Large Print. The Large Print version was available as an MS Word document and could be printed, though not directly from the menu.

[69] HFS Single Character Short.

[70] In the SVT1 study, this level was termed *cosmetic* rather than *very minor*.

[71] Phase 3 focus groups were numbered 4 through 8 consecutively by date.

[72] Individuals who are blind (as opposed to deaf-blind) also noticed a mismatch between what was voiced and what was displayed on the Braille display. It is not clear the extent to which this difficulty noted with JAWS for Windows would also be found in Window-Eyes. By some accounts, at the time that the research was conducted, Window-Eyes provided better support for refreshable Braille than did the more popular JAWS for Windows.

Also, a deaf-blind participant (#33) noted the use of conventional tables as a problem "because [in navigating through the table] there's no reference point for which column you're in. Say you're in the middle, row three, cell four, you don't know that that's where you are because there's nothing before or after it, or above or below it. So you don't really have a visualization of where you are in the table."

[73] ZoomText had also been an option for SVT3 but had not been requested. Among SVT3, VOS, and HFS, only HFS was expected to be useful (via screen readers) with Braille

displays. Notwithstanding his appreciation of HFS, he later expressed his preference for SVT3 over HFS for tests.

[74] The names shown here for the HFS and VOS systems reflect the names used in this report rather than the older, less meaningful names that were actually used.

[75] To avoid specifying the country of residence and to thereby preserve the anonymity of participants, this document uses the term *African North American* rather than more specific terms such as *African American*.

[76] Participants were alerted to the fact that even though the section directions page referred to three items in the section, SVT3 (i.e., SVT3 Short) actually had only one item per section.

[77] Actually called the "method-selection" tool during the study.

Appendix A

Features of Different Versions of the Self-Voicing Test

All versions of the ETS Self-Voicing Test system (SVT1, SVT2, and SVT3) had the following characteristics:

1. The system used the self-voicing approach (rather than the screen reader approach) for providing speech access to software applications. Participants were free to go back and forth between the question (stem), choices (options), and passage (where applicable).

2. The system provided keyboard access to the various commands.

3. The system provided auditory descriptions for visuals (Chisholm, Vanderheiden, & Jacobs, 1999). These auditory descriptions consisted of a prerecorded audio of a narrator describing the visual. Without auditory descriptions, the content provided uniquely by the visuals would be unavailable to individuals who are blind.

4. The visual display of the system used large white text on a black background, a combination that is often helpful to individuals who are partially sighted.

5. No specific time limit was placed on participants to complete the test.[59] (This feature is an incidental rather than essential part of the approach.)

6. The test taker could not go back to previous items. (This feature is an incidental rather than essential part of the approach.)

Enhancements in SVT3

The SVT3 used in this project also included several enhancements that distinguish it from the system used in the SVT2 field evaluation systems. The major enhancements are as follows.

1. The system was ported to a more stable software platform. In SVT2 it was not uncommon to have the user session interrupted by a system problem. In response, the system was ported to the ETS C3 delivery platform[60]—a Web-based test prototyping system that is used within research to deliver a wide range of prototypes—and certain routines for handling test-taker commands were reprogrammed. Employing a robust, standards-based system like C3 permits rendering multiple outputs from a common XML-based store, without modifying the primary source material. These changes

appeared to substantially reduce the frequency of system problems as well as allow use of the more sophisticated infrastructure for saving response data.

2. Auditory cues about progression through content were added. In SVT2, it was fairly common for a person to incorrectly assume that they had heard the entire passage when in fact there was more to be heard. To address this problem, after voicing a paragraph (in a passage or question [stem]) in which there were more paragraphs to hear, the system would say in a different (i.e., male) voice the word "More," meaning that there is more content to hear. Furthermore, after having voiced the final paragraph in a passage or question, the system would say "End," meaning that one is at the end of that component. (Generally, users found these auditory cues "very helpful.")

3. Test directions were made more extensive and put on their own pages. In SVT2, the test directions were audio-only and consisted basically of a few words of welcome plus the instruction to press Enter to continue or Escape to hear the message again. In SVT3, the test directions occupied their own 267-word page and provided concrete instruction on how to navigate through the page and how to obtain help regarding available commands.

4. Section directions were made more extensive. In SVT2, section directions appeared on a distinct page but typically consisted of a few dozen words. In SVT3, they are considerably longer, the longest being 318 words in length. (It should be noted that this longest set of section directions contained some general navigation information that could otherwise be placed in general test navigation instructions.)

5. Item directions were made more extensive and put on their own page. In SVT2, the item directions were audio only and typically consisted only of a few words. In SVT3, they occupy their own page and may contain a few dozen words.

6. The capability of reviewing section and item directions was added. In SVT2, one could not re-hear (or re-see) section or item directions. In SVT3, there were distinct commands to return to section or item direction information in a pop-up window. Having navigated through the pop-up window as desired, the test taker presses Enter to return to the item.

7. The system was modified to yield a reasonable result from the most basic command—"voice next paragraph" command (Control-Down Arrow). In SVT2, on any new page, be

it an item page or a directions page, Control-Down Arrow would have no result, often leaving test takers puzzled about what to do next. (They were generally supposed to first navigate to a component, e.g., Control-Shift-P for passage, Control-Shift-Q for question, Control-Shift-C for choice list, before using a command such as Control-Down Arrow.) This might have been remedied by prompting the test taker about the necessity of first navigating to a component (e.g., passage, question, choice list). However, we went a step further by actually placing the test taker at the beginning of the "first" component, that is, in the passage if there is a passage or in the question if there is no passage. With this arrangement, pressing Control-Shift-Down Arrow starts voicing the first paragraph of that first component. In order to avoid biasing the test taker to read the passage first instead of the question, the item directions would typically alert the test taker to the fact that although they are started in the passage, they are free to navigate to any component at any time.

8. The answer confirmation sequence was modified. In SVT2, the confirmation sequence was useful for the single-selection multiple-choice format but not for multiple-selection multiple-choice format. The test taker interacted with a choice list as follows.

    a. Navigate to the choice list using Control-Shift-C.

    b. Navigate from choice to choice using Control-Up Arrow or Control-Down Arrow.

    c. After hearing the desired choice, press Enter to select it. If the test taker selected choice 2 at step c, then the system voices, "You have selected number 2. Press Enter to go on, Escape to cancel." At the same time, the system displays a modal dialog box with the message "Are you satisfied with your choice?" along with an "OK" button and a "Cancel" button. The system does not allow the test taker to leave the item without selecting a choice.

    The confirmation sequence was modified by adding a step.

    a. Navigate to the choice list using Control-Shift-C.

    b. Navigate from choice to choice using Control-Up Arrow or Control-Down Arrow.

    c. Press the spacebar after hearing the desired choice in order to select it. If choice 2 had been selected and the choice was "He opened the door," then the system voices "You have selected choice 3: 'He opened the door.'" One can repeat the selection process for multiple-selection multiple-choice items by returning to step b.

d. Press Enter. If the test taker selected a choice at step c, then the system displays a modal dialog box with the message, then the system simply proceeds to the next page. On the other hand, if the test taker did not make a selection at step c, then the system displays a modal dialog box with the message, "You have not selected any choice. Press Enter to confirm that you want to skip this item" along with an "OK" button and a "Cancel" button.

The fact that the system voices the choice back to the test taker gives him or her the ability to check if the system has correctly registered their intention. Furthermore, adding the extra step to the confirmation sequence allows the possibility of another important response format—multiple-selection multiple-choice. A multiple-selection multiple-choice item was implemented in the system but was *not* part of SVT3 (i.e., SVT3 Short). The fact that SVT3 allows the test taker to skip items is incidental rather than essential to the approach. (Neither SVT1 nor SVT2 allowed the test taker to skip an item.[61])

9. Automatic voicing of the question was almost entirely eliminated. When one navigated to the *question (stem)* in SVT2, it was automatically voiced, whereas navigating to the *passage or choice list* required the test taker to issue a navigation command (e.g., Control-Down Arrow) to begin voicing the content. This inconsistency was eliminated by always requiring the test taker to issue a navigation command to begin voicing of test content. The only exception was the first item following a listening comprehension stimulus, which was voiced automatically to minimize the delay between hearing the stimulus and responding to the first item.

10. Greater use was made of a "phonetic alphabet"[62] for single letters in commands. In SVT2, the single letters found in most commands were often distinguished from each other. For example, the command "Control-Shift-D" might be mistaken for "Control-Shift-T" unless the former were stated with a word starting with "D" such as "Control-Shift-D as in Delta." This phonetic alphabet approach was adopted with much greater consistency in SVT3.

11. Directions pages were changed to use ordinary navigation and were no longer mandatory. In SVT2, each section had a directions page for which voicing was initiated with a special command (Control-Shift-D), and listening to the whole page was mandatory. In SVT3, reading the page was no longer forced, and navigation was activated as with other components by commands such as Control-Down Arrow.

12. Limited support for mouse was added. One could not operate SVT2 with a mouse but was required to use the keyboard, which was a disadvantage or inconvenience for users who preferred to use the mouse. SVT3 provided partial mouse support by allowing the user to select a choice with the mouse or by pressing the spacebar. (This feature was not used in the Phase 3 study.) For both SVT2 and SVT3, the dark interior of a circle became light-colored when a choice was selected.

13. The system was made available via an Internet connection. The studies for SVT1 and SVT2 involved running SVT entirely from a laptop. On the other hand, VOS—except for listening comprehension, which used large multimedia files—was capable of running on basically any multimedia-capable computer with a modem (55K) or Internet connection, a recent version of MS Internet Explorer, and the MS Speech API version 4.0. An Ethernet connection was sometimes adequate for short listening comprehension stimuli. HFS was capable of running over the Internet without the speech API software. Use of the Internet connection for the Phase 3 study was negligible.

Appendix B

Additional Detail Regarding the Phase 3 Evaluation

The material in this appendix is a supplement to the material in the main body of the report. In order to maintain some continuity in the description below, it contains a small amount of material that is duplicated in the main body of the report.

Method

Sample

The sample for the field test (Phase 3) consisted of 15 participants with blindness (BL, *n* = 4), low vision (LV, *n* = 2), learning disability (LD, *n* = 3), deafness (DF, *n* = 2), deaf-blindness (DB, *n* = 2), or no disability (ND, *n* = 2).[63] Recruited participants were a convenience sample. All students were adults and were either college-bound (*n* = 2) or already had experience with higher education. Of the 15 participants, 8 were nonnative speakers of English.[64] Of the 5 nonnative participants who were asked, 4 rated themselves as being able to speak "as fluently as an educated native speaker of English," while 1 (BL) considered himself at a lower speaking level ("can participate in most conversations").[65] None of the participants had taken the TOEFL test. The mean age was 34 years, and the ages ranged from the 20s to the 50s. Of the 15 participants, 5 were female and 10 were male. Other details of student background can be found in Appendix D.

Materials Preparation

Early plans and developments shaped the nature and variety of delivery and administration methods that were evaluated for this project. In addition to the three systems, the following other materials were developed for the evaluation.

Administrator Screens

Administration screens are intended to allow a test proctor to select a testing option from a list and then to invoke its delivery.

- *Regular administrator screen.* The "regular" administrator screen provided text input fields for the password and the student identifier,[66] a drop-down box to select 1 of 10 testing options,[67] a "Run Test" button to launch the test, and a "Reset" button to clear the text input boxes and to reset the drop-down box to its default value (option #1). Pilot testing on an early version of the regular administrator screen suggested that some users of screen reader software (who were generally blind) have difficulty filling out such forms and that another version may be needed.

- *Voiced administrator screen.* This presented an administrator screen, but with self-voicing so that a person with a visual disability could use it. The screen was simplified relative to the regular administration screen and contained only a list of 12 testing options (as shown in Appendix C). Upon arriving at the screen, a brief auditory message about how to navigate the screen is presented. Administrators navigate through the list of testing options using the Up Arrow and Down Arrow, and then press Enter to invoke the option.[68]

Testing-Selection Tool

This tool was a Microsoft Excel spreadsheet table for assisting an administrator to determine the most appropriate testing option. A key portion of this table is shown in Appendix F. The table has 12 important row headings, one for each of 12 testing options (e.g., different variants of SVT4, HFS, VOS). There are 9 important column headings, one for each of nine profiles, where each profile is characterized by a disability status and sometimes by communication preferences (1. Blind, Braille display; 2. Blind, Synthetic Speech, Braille not essential, etc.). In the cell defined by the intersection of a testing option and a profile is found a rating (Good, Marginal, N/A) that tells how well, in the estimation of a researcher, the testing option fits with the basic accessibility requirements of persons with that profile. The table could be used simply by inspection or its use could be facilitated by the "data filter" capability built into Microsoft Excel. The data filter capability allows one to quickly select those testing options (rows) that meet a certain criteria, e.g., "Good" fit to one or more profiles. It must be emphasized that the testing-selection tool considers only a few of the considerations that might need to bear upon actual operational test delivery decisions. For example, the tool does not specifically take into account aspects of a test delivery (e.g., read-aloud feature) and how that might threaten to invalidate measurement of a certain construct (e.g., reading comprehension). This testing-selection tool could be used visually or via a screen reader.

Large-Print and Raised Line Braille Drawings

A large-print hard-copy figure and a raised line drawing (with Braille labels) for the quantitative comparison item in the SVT3 (SVT3 Short) and HFS[69] was available. The GRE program provided the large-print and raised line drawing.

Procedures

This section focuses on the field test that occurred during Phase 3 of the project. The field test consisted of three parts: (a) individualized observation sessions, in which all participants were involved; (b) focus groups, to which all were invited (12 of the 15 invited participants were able to attend); and (c) tool review sessions, which involved 5 of the 15 invited participants in either individualized or group settings. In sessions involving participants who were deaf or deaf-blind, professional sign language interpreters were used to facilitate communication. Participants received monetary compensation for their participation.

Observation Sessions

Observation sessions for Phase 3 were held in four locations: ETS in Princeton, New Jersey; the Adaptive Technology Resource Centre (ATRC) at the University of Toronto, Ontario, Canada; Gallaudet University, Washington, D.C.; and the Helen Keller National Center for the Deaf-Blind (HKNC), Sands Point, New York. Observations sessions were set for three hours, though in some cases the session ran longer or part of a session needed to be conducted by phone or e-mail. The observation session consisted of five parts:

Welcome and orientation. Consent forms were reviewed and signed. Participants were told that they would be asked to help identify strengths and weaknesses of one or more systems. Participants were encouraged to do their best, although their performance would not affect their grades or job.

Background questionnaire. Personal interviews were administered before the participants used the systems. The participants were asked about their educational and professional background, disability, and language experience and proficiency, as well as their prior experience with and reliance on alterations such as speech synthesis technology, human readers, and Braille technology.

Assignment to a treatment set. Each participant was informed of the treatment set to which they had been assigned. (See main body of report.)

Cycles of instruction and observation. For each delivery option tried, there were periods of instruction and observation.

- Instruction. The instruction was intended to be minimal so that we could place greater reliance on the computer-based directions. Generally, almost no introductory instruction in the use of the systems was provided.
- Observation. Generally, participants were observed without interruption until they appeared to have a problem. At that point, they were instructed in the use of the severity rating system. The scale has five levels, four of which indicate user perception of a usability problem: 0 = I don't agree that this is a usability problem; 1 = Very minor [cosmetic] usability problem;[70] 2 = Minor usability problem; 3 = Major usability problem; 4 = Catastrophic usability problem, which must be fixed before the test can be used in important settings like college or graduate school admissions (see Nielsen, 1994). Generally, participants had as much time as they desired to answer the questions for the first method that they tried; nevertheless, if time was running short, they were encouraged to move along. Each participant was observed using the system by one and occasionally by two or three researchers during the field test. Participants were generally encouraged to try to solve navigation issues themselves. For example, in SVT3, they were encouraged to look for answers to navigation problems by looking at the command menu (Control-M), section directions (Control-Shift-E), or item directions (Control-Shift-E). Some participants were given instruction in additional commands. If it became clear that additional assistance was necessary, the researcher would help them know what to do next. During or after the difficulty, the researcher would discuss the problem with the participant to try to identify the nature of the usability problem, its possible cause, how serious of a problem it was (using the severity rating scale), and how the problem might be solved. Questions about usability were also asked via interview.

A few additional questions were administered in a post-session questionnaire following the entire observation session because they involved comparisons of two methods.

Focus group sessions. Five exploratory (Phase 3) focus group studies were conducted on August 23 and 29, and September 20, 26, and 28, 2001.[71] Two focus groups were conducted at the Adaptive Technology Resource Centre (ATRC) at University of Toronto, Canada; one at

ETS in Princeton, New Jersey; one at Gallaudet University in Washington, D.C.; and one at the Helen Keller National Center for Deaf-Blind Youths and Adults, in Sands Point, New York.

In the focus group sessions, participants were allowed to share their opinions regarding the main strengths and weaknesses of the system. They were also encouraged to discuss priorities for improvement. The principal investigator served as a moderator, and the sessions were videotaped.

The participants who reviewed the system in the observation sessions were invited to participate in a focus group. Of those 15 participants who reviewed the system, 12 took part in the focus group sessions.

Participants generally used their first names. Each focus group session was conducted in a 2-hour block and included approximately 1 hour and 40 minutes of discussion.

For the qualitative data analysis, the audio from the videotapes was transcribed. Since similar questions were asked in all five focus groups, the data were combined and analyzed as a unit.

Tool Review Sessions

Tool review sessions were conducted, during which time 5 participants reviewed an administrator screen (regular voicing) and/or a testing-selection tool (with or without screen reader and refreshable Braille). Criteria for selection of participants for tool review sessions included prior participation in a Phase 3 observation session and status as blind or nondisabled. Participants tended to have higher than average educational attainment. Among the 3 blind participants (#20, 31, 35) and 2 nondisabled participants (#30, 36), 2 had completed master's degrees, 2 had completed four-year degrees, and 1 had completed a two-year degree. Sessions typically lasted about 15 to 25 minutes. Participant interaction with the system was individualized, but for 3 of the participants (#20, 35, 36) that interaction occurred in the context of a single group meeting with the researcher. The specific nature of the tasks is described in the Findings section.

Results

Results From Observation Group Sessions

Severity Ratings for the Usability Problems

As noted earlier, the participants rated the severity of usability problems according to a scale that was closely adapted from Nielsen (1994). The scale has five levels:

0 = I don't agree that this is a usability problem.

1 = Very minor (cosmetic) usability problem.

2 = Minor usability problem.

3 = Major usability problem.

4 = Catastrophic usability problem, which must be fixed before the test can be used in important tests like tests of college or graduate school admissions.

Across all administrations of the three delivery systems, 172 problems (i.e., Levels 1 through 4) were cited (see Table B1), 38 of which were rated as "catastrophic" (see Table B2).

Table B1

All Usability Problems

	Disability status						
Delivery method	BL	LV	LD	DF	DB	ND	Grand total
HFS	11	4	10	7	11	19	62
VOS	n/a	n/a	n/a	5	n/a	n/a	5
SVT3	34	21	21	n/a	n/a	29	105
Grand total	45	25	31	12	11	48	172

Table B2

Usability Problems Rated as "Catastrophic"

	Disability status						
Delivery method	BL	LV	LD	DF	DB	ND	Grand total
HFS	4	0	2	3	7	4	20
VOS	n/a	n/a	n/a	2	n/a	n/a	2
SVT3	3	5	2	n/a	n/a	6	16
Grand total	7	5	4	5	7	10	38

Table B3 shows examples of catastrophic usability problems. These are tabulated by disability status and delivery method and also show language status and gender.

Table B3

Examples of Usability Problems Cited as "Catastrophic"

Disability	Method	ID	Description	Language	Gender
BL	SVT	35	The system did not provide a tutorial.	Nonnative	M
		31	The text description of the math visual is not precise. The phrase "lies above" is too vague.	Nonnative	F
		31	The system doesn't allow users to review items.	Nonnative	F
	HFS	31	The system lacks a set of tips for using a screen reader.	Nonnative	F
		31	Drop-down boxes, radio buttons, and checkboxes are very hard to use with a screen reader.	Nonnative	F
		20	One can enter text but not edit in Forms mode (using JAWS for Windows).	Native	M

(Table continues)

Table B3 (continued)

Disability	Method	ID	Description	Language	Gender
		20	Internal links do not operate reliably (often take user to top of page).	Native	M
DB	HFS	34	Braille display does not navigate properly, taking the user to the top of document rather than going down the page properly (using JAWS for Windows).	Nonnative	F
		33	System did not say how to control Internet Explorer (e.g., tab, down arrow).	Nonnative	M
		33	Lacked training (hour prior to test).	Nonnative	M
		33	Can't hold place or set bookmark and quickly return to it.	Nonnative	M
		33	Takes a lot of time.	Nonnative	M
		33	Hard to find information.	Nonnative	M
		33	Accessing HTML with MS Internet Explorer and Netscape Communicator is extremely difficult.	Nonnative	M
DF	HFS	38	It is not clear that the check box item has more than one possible choice. Need to alert user in bold.	Native	F
		37	Structure (organization) of item is unclear.	Nonnative	M
		37	Hated drop-down box response format.	Nonnative	M
	VOS	38	Cannot review items.	Native	F
		37	Next button failed (system bug).	Nonnative	M

(Table continues)

Table B3 (continued)

Disability	Method	ID	Description	Language	Gender
LD	SVT	39	Instructions for using Control-Shift-P (followed by Control-Down Arrow) to review the picture in the quantitative comparison item are not clear in the item directions. [a]	Native	M
		27	Had to restart test because of illegal operation during listening comprehension (system bug).	Native	M
	HFS	39	Not clear that more than one answer is possible for the items with checkboxes response type.	Native	M
		39	Lacked spell checker.	Native	M
LV	SVT	40	Lacked training/familiarization.	Native	M
		40	Accidental use of browser commands (e.g., Control-P instead of Control-Shift-P) confuses user (and requires proctor to fix).	Native	M
		40	Using "Control-" for some commands and "Control-Shift-" for others is confusing.	Native	M
		40	Picture needs white lines on black background rather than black lines on white background.	Native	M
		32	Accidental use of browser commands (e.g., Control-P instead of Control-Shift-P) confuses user.	Native	M
ND	SVT	36	System lacks scroll bar on command menu.	Nonnative	M
		36	The "time remaining" key does not work during directions.	Nonnative	M
		36	Section directions are not available from within item directions.	Nonnative	M

(Table continues)

Table B3 (continued)

Disability	Method	ID	Description	Language	Gender
		36	Pause/resume did not work reliably with problem-solving directions or in command menu.	Nonnative	M
		36	Lacks practice session.	Nonnative	M
		30	Could not navigate to passage. Control-Shift-P did not work.	Native	F
	HFS	30	Accidentally pressed Enter and caused to submit before completing test.	Native	F
		30	Using up/down arrows to move down screen caused radio button answers to change.	Native	F
		30	Failed to tell user to scroll down to proceed from one item to next.	Native	F
		30	Does not confirm that answer choice was accepted before proceeding to next item.	Native	F

[a] These instructions referred to here were intended to deal with the problem that, as the audio of the yellow word highlighting progressed, the picture at the top of the panel gradually scrolled off the top of the screen. Instructions were intended to tell how to review the picture.

Results From Focus Group Sessions

This section focuses on major issues encountered in the focus group sessions.

HFS

Navigation. One of the major issues that the focus group addressed regarding the HFS system was the navigation issue, particularly with screen readers. For example, a low-vision participant (#32) who had recently begun using a screen reader said, "I found it hard if I was looking for a particular point, if I was trying to do a search or things like that and it just wasn't working. Maybe if I had more time, but in the time I had it wasn't working very well in terms of finding specific points in there." Internal links from the top of the test document to various sections of the document were suggested as a way to improve navigation. Participant #38 (DF) suggested, "Since everything for the HFS was on one page, I thought it would be a good idea to link all the various things that are available in HFS, right on the top of the page. And sort of make an anchor to each different section, so you could navigate there right off."

Slowness of reloads. A blind participant using both screen and Braille display (#35) said, "[P]art of my problem with the HFS was of course that loading the thing took a while." This individual was the only individual to use the Window-Eyes screen reader. Alternating between reading mode (MSAA mode) and writing mode (forms mode), there was roughly an 8-second pause to load the page in Window-Eyes. This occurred at least once each time a question was answered.

Screen reader and Braille difficulties. There were problems of navigation when using screen reader and Braille display. Both deaf-blind participants used the screen reader and refreshable Braille display. One of these individuals (#34) expressed, "It was hard for me to find the top of a particular line, or the beginning of a particular line. And the top of the document was really hard for me to find … For example, when I tried to answer some of the questions, and I needed to go back to the top of the page so that I could refresh my memory, I had difficulty getting back to the spot that I needed." The other deaf-blind participant (#33) mentioned the difficulty that he had when using MS Internet Explorer and contrasted it with the relative ease of using MS Word: "When I tried to read the document, paragraph by paragraph, I would down arrow and it would jump down to the bottom of the page. And it would skip a few lines. So that was a point of concern for me, because I didn't want to miss any of the important information. If the document was in MS Word, it would scroll straight down line-by-line, and that would be fine. You didn't miss any of the information." He also affirmed that these problems were not unique to HFS but that they occurred whenever he accessed an HTML page with MS Internet Explorer. Regarding the mismatch between what a sighted hearing observer might see, he said, "The speech and the Braille display didn't match. The speech didn't match at all. It didn't match. … So, that's just the way it operates. So that when you're on the screen, it's always going to be a little bit different. You're never going to be simultaneous."

Another frustration was the difficulty of finding and editing text input fields relying on the Braille display. Both the deaf-blind and blind users of Braille noticed this problem. A deaf-blind participant (#34) said, "When I started to do like number one, for example, I was going to respond to that question, the blank didn't show up. It was like on the Braille there was nothing there. I was trying to respond, but there was nothing on the Braille display."[72] She noted, "The

Braille display is not recognizing the graphics or the radio buttons. But obviously, it's not working very well because it's not showing my answer up on the screen. And I also didn't see the multiple-choice ones: A, B, C, D. All that showed on the Braille display was the words, but not the A, the B, the C, the D."

Scrolling. The issue of scrolling appeared to be a concern. Several participants considered the scrolling in HFS somewhat distracting. A deaf participant (#38) suggested that scrolling be avoided if at all possible: "I could be totally wrong, but I think that deaf people tend to read things very fast because we depend on reading. Everything is visual. Any assistive device is reading. Our pagers that you see us using, our TTYs, our deaf telephones, our closed captioning on the television, everything is reading. So we are very fast at it; we don't want to scroll, we want to just see everything in one stop—one stop shopping. So [if] . . . the problem's on one side, the question's on the other, that allows me to skim that quickly. [I] give you the answer, and move on, rather than read this, go down, look at the answer and just come back up, you know, it's a waste of time. Just have it all right there . . . " She further elaborated that scrolling builds the anxiety for her.

Clarity of speech. A blind participant (#20) commented, "The 'pro' of the HFS was the clarity of the speech … I was familiar with the speech program that was on the computer, and I was able to, not necessarily as fast as I'd like to, but get to the information that I needed to get to. …"

Strengths of HFS. Perceived strengths of HFS included customizability and Braille use. A low-vision participant (#32) felt that HFS allowed him to customize the system a little more to his own liking. He said, "You know if you prefer the ZoomText or you prefer the JAWS [screen reader]."[73] Also, a blind participant (#20) said, "The fact that the HFS used JAWS and was able to simultaneously use a Braille display I found helpful."

SVT3

Voicing. The voicing aspect of SVT3 was considered a strength. For example, participant #27 (LD) said, "[I] enjoyed the presence of the voice. The fact that there was the voice presence is good." He also said, "I found the reading aloud enhanced my comprehension of the written test." However, the quality of the synthesized speech was cited by several individuals as needing improvement. (It had been similarly cited in previous studies.) A blind participant (#31) said, "Sometimes it's hard to catch certain words or, you know, sometimes certain letters, like Ps and

Bs and Cs and things like that, that can kind of get mixed up. I mean, evidently that must have been seen at some point, because it always says M for Mike."

Tracking. In SVT3, the yellow highlighting goes from word to word in synchronization with the speech.. A low-vision participant (#40) noted, "The word tracking was excellent. … If we could add to that feature, the yellow highlighting for the special needs community. … There is a huge demand for that because to me it helped because if you lose your spot, I tend to lose my place with my vision, and that is common for different types of special needs people. It keeps you on track."

Descriptions of the math figures. Although the text descriptions of math visuals were thought to be helpful, it was emphasized that their quality needed to be excellent, and the need for access to tactile drawings was emphasized. Participant #31 (blind, nonnative) said, "I would think that [what] I'd like to see in the next version is clear explanations of the math … If you just have an explanation, and you cannot see the figure, but you just have it explained in words, it's just not as clear as when you actually have, it's hard to visualize." An individual who is deaf-blind (nonnative, #33) indicated, "Sometimes, I'll read an explanation and I really don't understand it. But if I have a picture, like what you have, it's a tactual picture, then I can understand it better."

The need for tutorials. Several participants expressed a need for tutorial materials for getting used to the system. During the focus group discussions, some participants realized that some of the features that they thought were missing from the system were actually available in the system. For example, several participants suggested providing a command that allows changing the speed of the speech, and only then realized that the feature was already there. This issue might be addressed through improved training materials. A blind participant (#31) cited the need for: "running through how the program works and have examples to … go through… Because then, you know, as you get familiar with it, you could navigate much quicker rather than, … just starting out during the test."

Customization. A low-vision participant (#40) appealed for greater flexibility in SVT, "… Like whatever the system uses, you should be able to change. There has to be some sort of flexible function or … that changes background colors and that sort of thing, so a person could sort of personalize to one that's easier for them. … If I could set the colors up to the colors that I like, that would be ultimately the best. You know, if it provided for customization …"

One suggestion was to save each test taker's customization settings so that they could be retrieved later for immediate use. This would save time trying to figure out the best configuration settings for each new session. One mechanism for saving these settings would be with a "smart card." As one low-vision participant (#40) noted, "The voices I mean, either if it's voiced it'll help that person. Every special needs experience is specific …Special needs really have to be catered to that person. Meaning, the smart card concept would help that, so that it was configured for them better at the beginning, and that's for all of them, once again."

One blind participant (#20) who used both SVT3 and HFS commented: "I thought the method that was most effective was the Self-Voicing Test, the SVT. I thought it was easy to use and logical. … Where with using the HFS and using JAWS [for Windows screen reader for] navigating, though it was possible, was more difficult. I preferred the speech that JAWS uses over the SVT. SVT voice was adequate, but JAWS's voice, I thought, was clearer. And that it would be an ideal situation if an individual was able to use a Braille display simultaneously with the SVT. I was very impressed with the SVT."

Regarding disadvantages of SVT3, a blind participant (#35) said, "I think with SVT my features are verbosity and not being able to re-do your answer again. For me, I think since ETS does normally give you extra time for need, not being able to re-do your answer again might be more of an issue." This individual was quite concerned that SVT did not allow the test taker to go back to previous items: "Well, I'd definitely want to be able to review because if I am unsure, yeah, I would read the passage on my answer, so I might just really change my mind and want to change that answer." Participant #37 (DF) rated the inability to review items as a problem of "catastrophic" severity.

In general, a recurring theme was that the delivery systems need increased flexibility to allow individuals to interact with the systems according to their preferences.

Findings Regarding the Administrator Screens

In separate sessions, each participant using the *regular* administrator screen (participants 30 and 36, both nondisabled) was asked to enter the password and student ID, and then was directed to select a researcher-specified option in a drop-down box and then to invoke it by pressing Enter or click on the "Run Test" button. Each participant was given an introduction and orientation lasting about one minute, followed by use of the system for a minute or two. Both nondisabled participants indicated that the system was "easy" to use.

The *voicing* administrator screen is essentially a self-voicing pick-list with a brief auditory introduction. In separate sessions each participant (participants 31 and 35) was asked to navigate up and down the list and then to invoke a researcher-specified option by pressing Enter. Each participant was given an introduction and orientation lasting about 2 minutes, followed by use of the system for a minute or two. Both indicated that the system was "easy" to use.

Findings Regarding the Testing-Selection Tool

In separate sessions, each of 4 participants used the testing-selection tool. Each participant was asked to use the Excel spreadsheet "data filter" feature to show only those methods among 12 testing options (rows) that met criteria entered by the user in the header cell drop-down list for the column for the user profile of the prospective test taker. For example, 1 participant was asked to identify testing options that were rated as "good" for individuals with a profile called "Blind, SynSpeech, Braille not essential." This profile is intended to encompass individuals who are blind and rely on synthesized speech but for whom Braille access is not essential. By selecting "Good" in the filter for this profile, the user can show only those testing options that the researcher had previously rated as being good (i.e., potentially usable) by such individuals. In this case, the tool included a variety of SVT and HFS testing options but excluded the VOS and hard-copy print options.

In each case, the instruction in the purpose and operation of the tool lasted about 5 to 8 minutes and the user's interaction with the system lasted about 5 to 8 minutes.

Both nondisabled participants who used the testing-selection tool found it "easy" to operate. The 2 blind participants who used it with a screen reader found it somewhat difficult to use. One user indicated that she liked using Excel and used it as much as possible. One of the users indicated that it would be a useful tool for someone who was experienced in using Excel spreadsheets with a screen reader but that he did not have sufficient experience with Excel to make good use of it. Neither participant was able to use the screen reader to activate the drop-down list to operate the filters to show only those methods that would be "good" for individuals with a selected user profile.

Appendix C
Options in the Voicing Administrator Screen

1. SVT3 Basic (11 items)
2. SVT3 Reading Comprehension and Structure (5 items)
3. SVT3 Math (6 items)
4. SVT3 Short (5 items)
5. SVT3 Listening Comprehension (4 items)
6. HFS Basic (12 items)[74]
7. HFS Radio Button Short (5 items)
8. HFS Single Character Short (5 items)
9. VOS Complete (16 items)
10. VOS Writing (1 item)
11. Large Print (11 items)
12. Regular Print (11 items)

Appendix D

Background of Phase 3 Participants

Participant #20 is a blind, White, native-English-speaking male in his 40s. He holds a four-year college degree, and has received alterations such as extra testing time (1.5x), a human reader, private room, reader/scribe, Braille printer, portable note taker (Artic TransTYPE synthesizer/notetaker), Braille display, and Braille writer. He has had experience with speech synthesis technologies such as JFW and Kurzweil Reader. He used SVT3 (9 minutes) and HFS with JFW and refreshable Braille display (86 minutes).

Participant #27 is a White, native-English-speaking male in his 20s with a learning disability. He holds a four-year college degree, and has received alterations such as extra testing time (1.5x). He has no experience with speech synthesis technology. He used SVT3 (40 minutes) and HFS (15 minutes).

Participant #28 is a blind, White male in his 50s whose primary language is Greek. He holds a two-year college degree, and has received alterations such as modification of font size; Closed Circuit TV (CCTV), with which he is very comfortable; mechanical Braille note taker, with which he is very comfortable; and hard-copy Braille, with which he is very comfortable. He has no experience with speech synthesis technology. He used SVT3 (72 minutes).

Participant #29 is a White, native-English-speaking female in her 30s with a learning disability. She holds a four-year college degree, and has received alterations such as extra testing time (1.2x), with which she is very comfortable, and a quiet room to avoid distractions, with which she is also very comfortable. She has not had any experience with speech synthesis technology. She used SVT3 (85 minutes) and HFS (29 minutes).

Participant #30 is a nondisabled, White, native-English-speaking female in her 30s. She is currently enrolled in a Ph.D. program, and has never received any alterations nor had any experience with speech synthesis technology. She used SVT3 (81 minutes), VOS (8 minutes), and HFS (73 minutes).

Participant #31 is a blind, White female in her 20s whose primary language is French. She holds a two-year college degree, and has received alterations such as extra testing time (1.3x), with which she is very comfortable; oral delivery of essays, with which she is very comfortable; audiocassette, with which she is very comfortable; mechanical Braille note taker (Perkins Brailler), with which she is very comfortable; electronic Braille note taker, with which she is somewhat comfortable; refreshable Braille display as part of an electronic note taker, with which she is very comfortable; refreshable Braille with screen reader software, with which she is very comfortable; hard-copy Braille, with which she is very comfortable; and a quiet room to avoid distractions, with which she is somewhat comfortable. She has had experience with the speech synthesis technologies JFW, with which she is very comfortable, and pwWebSpeak (a self-voicing Web browser), with which she is somewhat comfortable. She used SVT3 (84 minutes) and HFS with JFW and refreshable Braille display (150 minutes).

Participant #32 is an African North American,[75] native-English-speaking male in his 50s with low vision. He holds a four-year college degree, and has received alterations such as screen magnification software, with which he is very uncomfortable; large print, with which he is very uncomfortable; modification of font size, with which he is very uncomfortable; modification of text or background color, with which he is somewhat comfortable; closed-captioned television (CCTV), with which he is very uncomfortable; audiocassette, with which he is very comfortable; and a human reader, with which he is very comfortable. He has experience with speech synthesis technologies such as JFW, with which he is somewhat comfortable; ZoomText Level 2, with which he is very comfortable; and Kurzweil Reader, with which he is uncomfortable. He used SVT3 (75 minutes) and HFS with JFW (21 minutes).

Participant #33 is a deaf-blind male in his 30s whose primary language is Bengali. He holds a four-year college degree, and has received alterations such as extra testing time (2-3x to unlimited), mechanical and electronic Braille note takers, refreshable Braille display as part of an electronic note taker, refreshable Braille with screen reader software, with which he is very comfortable; and hard-copy Braille, with which he is very comfortable. His experience with speech synthesis technology includes JFW, with which he is somewhat comfortable; Window-Eyes, with which he is somewhat comfortable; and OutSPOKEN (screen reader), with which he is very comfortable. He used HFS with Window-Eyes and refreshable Braille display (71 minutes).

Participant #34 is a deaf-blind, Hispanic female in her 20s. Her native language is Spanish, and she holds a high school degree. Alterations she has received include extra time, large print, modification of text color or background with CCTV, mechanical Braille note taker, electronic Braille note taker, refreshable Braille with screen reader software, hard-copy Braille, and a tactile interpreter. She has had experience with the speech synthesis technology JFW. She used HFS with JFW and refreshable Braille display (95 minutes).

Participant #35 is a blind, White male in his 30s. His primary language is Maratchi, and he is currently in a Ph.D. program. Alterations he has received include extra testing time (1.25x), with which he is somewhat comfortable; audiocassette, with which he is very comfortable; hard-copy Braille, with which he is somewhat comfortable; human reader, with which he is very comfortable; scribe/amanuensis, with which he is very comfortable; and a quiet room to avoid distractions, with which he is very comfortable. He has experience with speech synthesis technologies, including Window-Eyes, with which he is very comfortable; OutSPOKEN, with which he is somewhat comfortable; and Kurzweil Reader, with which he is very comfortable. He used SVT3 (58 minutes) and HFS with Window-Eyes and refreshable Braille display (16 minutes).

Participant #36 is a nondisabled Asian male in his 30s. His primary language is Farsi, and he is currently in a master's program. He has not received any alterations nor had any experience with speech synthesis technology. He used SVT3 (152 minutes) and HFS (23 minutes).

Participant #37 is a deaf male in his 20s whose primary language is Dutch. He holds a four-year college degree and has received some education conducted in American Sign Language. He has not had any experience with speech synthesis technology. He used VOS (16 minutes) and HFS (5 minutes).

Participant #38 is a deaf, White, native-English-speaking female in her 20s. She holds a four-year college degree and has received alterations such as a sign language interpreter. She has not had any experience with speech synthesis technology. She used VOS (19 minutes) and HFS (6 minutes).

Participant #39 is an African North American, native-English-speaking male in his 40s with a learning disability. He holds a four-year college degree and has received alterations such as extra time (2x) and a scribe/amanuensis. He has experience with speech synthesis technology (HelpRead). He used SVT3 (120 minutes) and HFS (9 minutes).

Participant #40 is a male in his 20s with low vision. His native language is Arabic, he holds a master's degree, and he has received alterations such as extra time (1.5x), modification of font size and text or background color, CCTV, audiocassette, human reader, human note taker, dictation software, and a quiet room to avoid distractions. He has experience with speech synthesis technology and voice recognition software. He used SVT3 (57 minutes) and HFS (20 minutes).

Appendix E

Directions for SVT3

This appendix provides the text for the test directions for SVT3 as well as the section directions and the item direction for the reading comprehension item.

Test Directions

You are beginning the E T S Self-Voicing Test Prototype.

This test provides navigation keys to allow you to read (that is, to voice) the content. At any time you can view a menu of commands by pressing Control-M (as in "Mike"). One of the most important navigation commands is Control-Down Arrow, which is made by holding the Control key down and then pressing and releasing the Down Arrow key. Pressing Control-Down Arrow causes the next paragraph to be read aloud. The system lets you know that there is another paragraph to be read by saying the word "More" at the end of the current paragraph. Press Control-Down Arrow now to advance to the next paragraph.

The word "More" at the end of the paragraph tells you there are more paragraphs to read in a given component. Examples of components include: a passage, a question, a choice list, or a direction page.

You can interrupt the voicing at any time with navigation commands. For example, if the system is voicing the current paragraph, pressing Control-Shift-Right Arrow causes the system to immediately begin speaking the next sentence. You can stop or halt the voicing at any time by pressing Escape. If you don't know what to do next, try pressing Control-Down Arrow (to move to the next paragraph) or Control-M (as in "Mike") for a list of commands.

Each of the items allows you to change your answer, but only while you are at that item. Once you leave an item, you can neither review it nor change your answer. Press Enter when you are ready to continue.

Reading Comprehension Section Directions

Section Directions for Reading Comprehension

The three[76] Reading Comprehension items in this section all use the same passage or portions of that passage. Each Reading Comprehension item has three major parts: a passage, a "question" (such as the words: "What is the main idea of the passage?"), and a choice list (that is, the list of possible answers). You can navigate to these three parts as follows: Press Control-Shift-P (as in "Passage") to move to the passage; press Control-Shift-Q (as in "Question") to move to the question; press Control-Shift-C (as in "Choice") to navigate to the choice list. Once you have navigated to one of these three parts, you can begin reading it paragraph by paragraph (or choice by choice) by pressing Control-Down Arrow. Similarly, Control-Up Arrow moves you to the previous paragraph or choice. When you are within the choice list and find the choice that you think is correct, press the spacebar to select the choice. If you change your mind, press spacebar again to de-select the choice. When you have selected your choice, and are ready to continue to the next item, press Enter.

Note you can also use the mouse to select or de-select a choice by clicking on the oval. When you first move to the item page, the system starts you in the passage, but if you like, you can navigate immediately to the question using Control-Shift-Q or to the choice list using Control-Shift-C. Other navigation commands allow you to: navigate sentence by sentence (Control-Shift-Left Arrow or Control-Shift-Right Arrow), word by word (Control-Right Arrow or Control-Left Arrow), or to have words spelled out (Right or Left Arrow within word-by-word navigation).

In addition, Control-Shift-I (as in "Item") allows you to review item directions and Control-Shift-E (as in "Echo") allows you to review section directions.

You may wish to press Control-M (as in "Mike") to review the various navigation commands. Press Enter to continue.

Item Directions for the Reading Comprehension Item

Item Directions for Reading Comprehension Item 1

Item 1 refers to a reading passage. At any time, you can navigate to the passage (using Control-Shift-P as in "Passage"), the item (using Control-Shift-Q as in "Question"), or choice list (using Control-Shift-C as in "Choice").

The test item starts you in the passage. Press Enter to continue.

Appendix F

The Testing-Selection Tool

Table F1 depicts the testing selection tool,[77] which encompasses the portion of a Microsoft Excel spreadsheet that allows the user to quickly identify which of several testing options were believed (by a researcher) to be most suitable to individuals with nine different profiles. Each profile is characterized by a disability status and possibly a preferred communication method. To aid presentation of the spreadsheet's content in Table F1, the description of the testing options in the spreadsheet is presented in the list below, rather than in the table.

1. SVT3 Basic (11 items). Includes Reading Comprehension, Structure, Quantitative Comparison and Problem Solving, and Data Interpretation (no Writing). Provides synthesized speech without relying on screen reader software. ZoomText is optional. May be useful to individuals with any of the following characteristics: blind, low vision, learning disability.
2. SVT3 Math (6 items). Includes Quantitative Comparison Problem Solving, and Data Interpretation. Provides synthesized speech without relying on screen reader software. May be useful to individuals with the following characteristics: blind, low vision, learning disability.
3. SVT3 Math (6 items). Includes Quantitative Comparison Problem Solving, and Data Interpretation. Provides synthesized speech without relying on screen reader software. May be useful to individuals with the following characteristics: blind, low vision, learning disability.
4. SVT3 Short (5 items). Includes one item from each of the following areas: Reading Comprehension, Structure, Listening, Quantitative Comparison and Problem Solving, and Data Interpretation. (No Writing.) Provides synthesized speech without relying on screen reader software. ZoomText is optional. May be useful to individuals with any of the following characteristics: blind, low vision, learning disability.
5. SVT3 Listening Comprehension (4 items). Provides synthesized speech without relying on screen reader software. ZoomText is optional. May be useful to individuals with any of the following characteristics: blind, low vision, learning disability.

6. HFSb Basic (12 items). Includes Reading Comprehension, Structure, Writing, Quantitative Comparison and Problem Solving, and Data Interpretation. (No Listening.) May be used with a screen reader/Braille display and/or ZoomText. May be useful to individuals with any of the following characteristics: deaf-blind (with Braille display), blind, deaf. Uses radio buttons for multiple-choice items.
7. HFS Radio Button Short (5 items). Includes one item in each of the following areas: Reading Comprehension, Structure, Writing, Quantitative Comparison and Problem Solving, and Data Interpretation. (No Listening.) Relies on radio buttons for multiple-choice items. May be used with a screen reader/Braille display and/or ZoomText. May be useful to individuals with any of the following characteristics: deaf-blind (with Braille display), blind, deaf.
8. HFS Single Character Short (5 items). Includes one item in each of the following areas: Reading Comprehension, Structure, Writing, Quantitative Comparison and Problem Solving, and Data Interpretation. (No Listening.) Relies on single-character input boxes for multiple-choice items. May be used with a screen reader/Braille display and/or ZoomText. May be useful to individuals with the following characteristics: deaf-blind (with Braille display), blind, deaf.
9. VOSb Complete (16 items). Includes Reading Comprehension, Structure, Writing, Listening, Quantitative Comparison and Problem Solving, and Data Interpretation. May be used with ZoomText. Individuals requiring use of speech output (blind, etc.) must use screen reader technology; however, high levels of expertise in the use of that technology in Web applications are necessary to gain value. May be most appropriate for individuals who are deaf or nondisabled.
10. VOS Writing (1 item). May be used with ZoomText. Individuals requiring use of speech output (blind, etc.) must use screen reader technology; however, high levels of expertise in the use of that technology in Web applications are necessary to gain value. May be most appropriate for individuals who are deaf or nondisabled.
11. Large Print (11 items). Requires printing from MS Word file. May be useful for individuals with low vision.
12. Regular Print (11 items). Requires printing from MS Word file. May be useful for individuals who are nondisabled and many others with disabilities.

Table F1

Testing Options and Preferred Communications Methods

Description of testing option [a]	Blind, Braille display	Blind, syn-speech, [b] Braille not essential	Low vision	Deaf	Deaf-blind, visual enlargement	Deaf-blind, Braille display	LD, self-voicing	LD, quick navigation, no syn-speech	Nondisabled
1.	No	Good	Good	N/A	Marginal	N/A	Good	N/A	Marginal
2.	No	Good	Good	N/A	Marginal	N/A	Good	N/A	Marginal
3.	No	Good	Good	N/A	Marginal	N/A	Good	N/A	Marginal
4.	No	Good	Good	N/A	Marginal	N/A	Good	N/A	Marginal
5.	No	Good	Good	N/A	N/A	N/A	Good	Marginal	Marginal
6.	Good	Good	Good	Good	Good	Good	N/A	Good	Good
7.	Good	Good	Good	Good	Good	Marginal	Good	Good	Good
8.	Good	Good	Good	Good	Good	Marginal	Good	Good	Good
9.	Marginal	Marginal	Good	Good	N/A	N/A	N/A	Good	Good
10.	Marginal	Marginal	Good	Good	Good	Good	N/A	Good	Good
11.	N/A	N/A	Good	N/A	Marginal	N/A	N/A	Marginal	Marginal
12.	N/A	N/A	N/A	Good	N/A	N/A	N/A	Good	Good

Note. The names shown here for the HFS and VOS systems reflect the names used in this report rather than the older, less meaningful names that were actually used. [a] These descriptions are detailed in the list above. [b] The term *SynSpeech* is an abbreviation for *Synthesized Speech.*

Appendix G

Sample XML Code

Following is a simple example of what XML markup might look like if it needed to represent multiple layers of item content that can be rendered via various output modalities:

```
<TestItem CLASS="ssmc" METHOD="list" TYPE="PS1">
<Stimulus TYPE="equation">
    <content MODE="standard">
        <Paragraph>
            <Sentence>2[2x + (3x + 5x)]
<MathSymbol>-</MathSymbol> (3x + 5x) = </Sentence>
        </Paragraph>
        <supplement>
            <Paragraph>
                <Sentence>This version of the
math equation is not intended for use with a screen reader.</Sentence>
            </Paragraph>
        </supplement>
    </content>
    <content MODE="mathML">
        <math DISPLAYSTYLE="true">
            <mrow>
                <mtext>2<mo
STRETCHY="false">[</mo>2x <mo>+</mo> <mo
STRETCHY="false">(</mo>3x <mo>+</mo> 5x<mo
STRETCHY="false">)</mo><mo STRETCHY="false">]</mo> <mo>-</mo>
<mo STRETCHY="false">(</mo>3x <mo>+</mo> 5x<mo
STRETCHY="false">)</mo> <mo>=</mo></mtext>
            </mrow>
        </math>
    </content>
    <content MODE="voicing">
        <Paragraph>
            <Sentence>The algebraic expression
reads, from left to right, as: two times open bracket two x
plus open parenthesis, three x plus five x close parenthesis,
close bracket, minus open parenthesis, three x plus five x
close parenthesis equals.</Sentence>
        </Paragraph>
    </content>
</Stimulus>
<Stem TYPE="prompt">
    <content MODE="standard">
        <Paragraph>
            <Sentence>Select the appropriate
solution from the list below.</Sentence>
        </Paragraph>
    </content>
    <content MODE="voicing">
        <Paragraph>
            <Sentence>Select the appropriate
solution from the following list.</Sentence>
        </Paragraph>
    </content>
    <content MODE="braille">
        <Paragraph>
            <Sentence>Choose the most appropriate
solution.</Sentence>
        </Paragraph>
    </content>
</Stem>
<Distractor_list>
```

```
    <Distractor>
        <Paragraph><Sentence>4x</Sentence></Paragraph>
    </Distractor>
    <Distractor>
        <Paragraph><Sentence>8x</Sentence></Paragraph>
    </Distractor>
    <Distractor>
        <Paragraph><Sentence>10x</Sentence></Paragraph>
    </Distractor>
    <Distractor>
        <Paragraph><Sentence>12x</Sentence></Paragraph>
    </Distractor>
    <Distractor>
        <Paragraph><Sentence>22x</Sentence></Paragraph>
    </Distractor>
</Distractor_list>
<Key>4</Key>
</TestItem>
```

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

Email: toefl@ets.org

Web site: www.ets.org/toefl

* America Samoa, Guam, Puerto Rico, and US Virgin Islands