NIST Special Publication 500-256

Improving the Usability and Accessibility of Voting Systems and Products

Sharon J. Laskowski
Marguerite Autry
John Cugini
William Killam
James Yen

NIST National Institute of Standards and Technology • Technology Administration • U.S. Department of Commerce

AUTHENTICATED
U.S. GOVERNMENT
INFORMATION
GPO

NIST Special Publication 500-256

Improving the Usability and Accessibility of Voting Systems and Products

Sharon J. Laskowski
Information Access Division
Information Technology Laboratory

Marguerite Autry
Senior Human Factors Engineer
User-Centered Design, Inc.

John Cugini
Cugini Consultants

William Killam
President and Principle Human Factors Engineer
User-Centered Design, Inc.

James Yen
Statistical Engineering Division
Information Technology Laboratory

May 2004



U.S. Department of Commerce
Donald L. Evans, Secretary

Technology Administration
Phillip J. Bond, Under Secretary for Technology

National Institute of Standards and Technology
Arden L. Bement, Jr., Director

National Institute of Standards and Technology Special Publication 500-256
Natl. Inst. Stand. Technol. Spec. Publ. 500-256, 97 pages (May2004)
CODEN: NSPUE2

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 2002

For sale by the Superintendent of Documents, U.S. Government Printing Office
Internet: bookstore.gpo.gov — Phone: (202) 512-1800 — Fax: (202) 512-2250
Mail: Stop SSOP, Washington, DC 20402-0001

Disclaimer

Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by NIST nor does it imply that the products mentioned are necessarily the best available for the purpose.

Note

This document was originally produced by NIST as a report to Congress to meet the requirements of Section 243 of the Help America Vote Act of 2002, Public Law 107-252. NIST first submitted the report to the U.S. Election Assistance Commission for review. On April 30, 2004, the Commission submitted the report to Congress.

Executive Summary

In the Help America Vote Act (HAVA) of 2002, Public Law 107-252, the Election Assistance Commission is mandated to submit a report on human factors, usability, and accessibility to Congress. Specifically, "…the Commission, in consultation with the Director of the National Institute of Standards and Technology, shall submit a report to Congress which assesses the areas of human factor research, including usability engineering and human-computer and human-machine interaction, which feasibly could be applied to voting products and systems design to ensure the usability and accuracy of voting products and systems, including methods to improve access for individuals with disabilities (including blindness) and individuals with limited proficiency in the English language and to reduce voter error and the number of spoiled ballots in elections."

This report was written to address this mandate. It describes the results of our review and analysis of related research, standards, guidelines, and evaluation methodologies. It also presents our assessment of their applicability to voting systems and products and to the process of qualification and certification testing. As a result of this investigation, we have compiled a set of recommendations that, if followed, should measurably improve the usability and accessibility of voting products and systems.

These recommendations are:

1) Develop voting system standards for usability that are performance-based, high-level (i.e., relatively independent of the technology), and specific (i.e., precise).

2) Specify the complete set of user-related functional requirements for voting products in the voting system standards.

3) Avoid low-level design specifications and very general specifications for usability. Only those product design requirements that have been validated as necessary to ensure usability should be included as "shall" statements in standards.

4) Build a foundation of applied research for voting systems and products to support the development of usability and accessibility standards.

5) To address the removal of barriers to accessibility, the requirements developed by the Access Board, the current VSS (Voting System Standards), and the draft IEEE (Institute of Electrical and Electronics Engineers) standards should be reviewed, tested, and tailored to voting systems and then considered for adoption as updated VSS standards. The feasibility of addressing both self-contained, closed products and open architecture products should also be considered.

6) Develop ballot design guidelines based on the most recent research and experience of the visual design communities, specifically for use by election officials and in ballot design software.

7) Develop a set of guidelines for facility and equipment layout; develop a set of design and usability testing guidelines for vendor- and state-supplied documentation and training materials.

8) Encourage vendors to incorporate a user-centered design approach into their product design and development cycles including formative (diagnostic) usability testing as part of product development.

9) Develop a uniform set of procedures for testing the conformance of voting products against the applicable accessibility requirements.

10) Develop a valid, reliable, repeatable, and reproducible process for usability conformance testing of voting products against the standards described in recommendation 1) with agreed upon usability pass/fail requirements.

In general, the **single most critical need** identified in this report is a set of usability standards for voting systems that are performance-based and support objective measures and associated conformance test procedures that can be used for the certification and qualification of voting products and systems. Usability, as we have defined it, is measured across all voters, including people with disabilities.

The report provides a plan for the development of the standards and test procedures that ensure usability. Research is necessary to validate our assumptions and initial conclusions and to make specific detailed recommendations for the tests. We recognize, however, that some States are facing procurement decision deadlines for products for upcoming elections and that they want to make wise choices that include usability and accessibility factors. In addition, all States are required to address HAVA voting equipment requirements by 2006. Therefore, we have also included some advice for informal usability and accessibility evaluation that can bridge the time gap between these deadlines and the development of new voting system standards for usability and accessibility.

Table of Contents

DISCLAIMER i

EXECUTIVE SUMMARY iii

TABLE OF CONTENTS v

1 INTRODUCTION 1

1.1 Scope 1

1.2 Background 2

1.3 Brief History of Standards and Testing for Voting Systems 3

2 BASIC TERMINOLOGY AND CONCEPTS 5

2.1 Definition of a "System" 5

2.2 Definitions of Accessibility and Usability 6
2.2.1 Disability 6
2.2.2 Accessibility 7
2.2.3 Usability and Usability Testing 8
2.2.4 Self-Contained, Closed, Accessible Products 8
2.2.5 Accessibility versus Usability 9
2.2.6 Usability in Practice (an Example) 9

2.3 Product Requirements, Usability, and Testing Methods 10
2.3.1 Type 10
2.3.2 Human Interaction 11
2.3.3 Levels 13
2.3.4 Specificity 13

2.4 Standards and Conformance Testing 14
2.4.1 Terminology of Standards 14
2.4.2 Pragmatic Issues for the Application of Standards 14

3 USABILITY AND ACCESSIBILITY REQUIREMENTS OF VOTING SYSTEMS 17

3.1 Implementation Examples of Functional Requirements for Voting 18
3.1.1 Implementation Variations 18
3.1.2 Example of Voting Product Design Variations 19

3.2 Potential Usability Problems in Voting Products 21
3.2.1 Usability Problems Prior to Success 21
3.2.2 Usability Problems Leading to Partial Failure 22
3.2.3 Usability Problems Leading to Total Failure 23
3.2.4 Examples of Potential Usability Problems 23

3.3 Potential Accessibility Problems in Voting Products 25

4 CURRENT USABILITY AND ACCESSIBILITY RELATED STANDARDS 29

4.1 Current (and Proposed) Voting Systems Standards related to Human Factors, Usability, and Accessibility 29
4.1.1 Requirements of HAVA 29
4.1.2 Current FEC Process: the VSS 30
4.1.3 IEEE Effort 36

4.2 Generic Usability and Accessibility Standards 37
4.2.1 Section 508 of the Rehabilitation Act of 1973, as Amended in 1998 38
4.2.2 ISO 9241: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) 39
4.2.3 ISO 13407: Human-centered Design Processes for Interactive Systems 40
4.2.4 ISO 16982: Ergonomics of Human-System interaction -- Usability Methods Supporting Human-Centered Design 40
4.2.5 ISO 10075: Ergonomic principles related to mental workload 40
4.2.6 ANSI/INCITS 354-2001: Common Industry Format (CIF) for Usability Test Reports 41

5 CURRENT HUMAN FACTORS ENGINEERING, USABILITY, AND ACCESSIBILITY RESEARCH 43

5.1 Background: Basic Research 43

5.2 Usability Research Related to the Design and Testing Process 44

5.3 Background: Applied Research 44

5.4 Usability Research Specific to Existing Voting Products 45

6 RECOMMENDATIONS 47

6.1 Overall Goal: Develop Measurable, Performance-Based Standards 47
6.1.1 Recommendation 47
6.1.2 Rationale 47

6.2 Specify Functional Requirements 48
6.2.1 Recommendation 48
6.2.2 Rationale 49

6.3 Avoid Detailed Product Design Specifications for Usability 49
6.3.1 Recommendation 49
6.3.2 Rationale 49

6.4 Address the Lack of Specific Research on Usability and Accessibility for Voting Systems on Which to Base Requirements 51
6.4.1 Recommendation 51
6.4.2 Rationale 52

6.5 Develop Design Specifications for Accessibility 52
6.5.1 Recommendation 52
6.5.2 Rationale 52

6.6 Develop Ballot Design Guidance 53

6.6.1 Recommendation 53
6.6.2 Rationale 53

6.7 Develop Facility and Equipment Layout Guidance 54
6.7.1 Recommendation 54
6.7.2 Rationale 54

6.8 Encourage Vendors to use a User-Centered Design Process 55
6.8.1 Recommendation 55
6.8.2 Rationale 55

6.9 Create Test Procedures for Accessibility 55
6.9.1 Recommendation 55
6.9.2 Rationale 55

6.10 Create Test Procedures for Usability 56
6.10.1 Recommendation 56
6.10.2 Rationale 56

7 ROADMAP FOR IMPLEMENTING THE RECOMMENDATIONS 59

7.1 Proposed Timeline 59

7.2 Short-Term 59

7.3 Long-Term Plans – 1-4 Years 60

7.4 Coordination with the TGDC 61

7.5 Proposed Next Steps for Testing and Standards Development 61
7.5.1 Proposed Testing 61
7.5.2 Proposed Standards Development 62

REFERENCES 63

APPENDIX A – GLOSSARY 71

APPENDIX B – DEVELOPING AND CONDUCTING USABILITY CONFORMANCE TESTING PROCEDURES 77

B.1 Test Environment 77

B.2 Voter Subsystem Testing 77

B.3 Poll Worker Subsystem Testing 78

B.4 Full System Testing 78

B.5 Standard Test Materials 78

B.6 Feasibility and Limitations 78

APPENDIX C – STATISTICAL DATA ANALYSIS **81**

APPENDIX D – REPORT METHODOLOGY **85**

APPENDIX E – AUTHOR BIOGRAPHIES **89**
Dr. Sharon Laskowski 89
Dr. Marguerite Autry 90
John Cugini 90
Bill Killam 91
Dr. James Yen 92

1 Introduction

The goal of this report is to describe how research and best practices from the human factors, human-machine and human-computer interaction, and usability engineering disciplines can be brought to bear to improve the usability and accessibility of voting products and systems. A major contribution of the report is a set of ten recommendations for developing standards, accompanying test methods, and guidelines that can measurably improve levels of usability and accessibility.

After the introduction, we discuss our assumptions and the information we used to generate the recommendations. We describe the current status of and testing process for voting systems, present an overview of the concepts of usability and accessibility, and discuss some related standards. We then present a detailed discussion of approaches that can be applied to improve usability and accessibility based on our review of relevant standards, guidelines, and testing and evaluation methodologies. We conclude with the set of recommendations and a discussion of short and long term activities that can help achieve the recommendations.[1]

1.1 Scope

The scope of this report is limited to human factors issues, that is, we are concerned with the process of the voter casting a ballot as intended and, to a lesser extent, the interaction of the poll worker with the voting system. This primarily involves the "user interface" the voter is presented by the system and the environment at the polling place. We have NOT examined issues concerning what happens after the voter casts a ballot such as the accuracy of counting the votes, the quality of the hardware and software, or the security of voting systems as these, in general, do not involve user interaction. Any approaches addressing these issues that do involve voter or election official interaction would require some analysis of the human factors and these should be addressed in future work.

Our analysis addresses issues pertaining to both voting products and voting systems. A **voting product**, as defined here, refers to a product procured from a vendor such as a Direct Recording Electronic **(**DRE) terminal.[2] By **voting system** we mean the combination of physical environment, voting product, ballot,

[1] Note to the Reader: If you wish, you may skim the recommendations in Section 6 before reading through the technical details of the report. This will put the detailed definitions and technical explanations for standards, testing approaches, research, and best practices that lead up to the recommendations into the proper perspective. However, to understand the rationale for the recommendations it is necessary to read through the technical details.

[2] The reader should note that the Glossary in Appendix A provided at the end of the report contains the definitions of voting and usability terminology used herein.

voter, and other persons involved in the voting process (e.g., poll workers and other election officials).

The bulk of the discussion focuses on the usability and accessibility of voting products for the voter. However, we also include usability issues pertaining to ballot design, the influence of the environment on accessibility as well as usability, and the setup and operation of voting systems by poll workers and election administrators. Further, we have constructed our recommendations for improvements so that they will fit into the existing and future qualification and certification testing frameworks for voting systems.

Note that we expect that these recommendations will be taken into consideration by the Technical Guidelines Development Committee (TGDC) when it becomes operational under the Election Assistance Commission (EAC) as described in the HAVA.

1.2 Background

There are many examples, some highly publicized, of voter confusion possibly caused by usability and accessibility problems (McIntire, 2003; Caltech–MIT 2001). Bedersen et. al identified a number of potential usability problems with DRE's to be used in Maryland (Bederson & Herrnson, 2002). Susan King Roth's 1998 article pointed out problems with readability, legibility, organization, and height (Roth, 1998).

A 2002 report (Burton & Uslan, 2002) from the American Foundation for the Blind's AccessWorld describing informal testing by 15 blind and low vision users reported "tremendous improvements over the way in which people who are blind and visually impaired currently vote" but also stated there was "certainly room for improvement." The report even cited one machine as preferable since it had "a lesser tendency to cause confusion." Additional informal testing at the National Federation of the Blind (NFB) has shown a number of accessibility or usability issues associated with nearly all of the six modern DRE devices they tested. Also, it should be noted that both the AccessWorld and NFB studies were performed on voting products with features specifically designed for voters with disabilities.

It also appears that the problems of voting product usability and accessibility are not felt equally across the voter population. The U.S. Civil Rights Commission reported in (Voting Irregularities, 2001) that "Poorer counties, particularly those with large minority populations, were more likely to use voting systems with higher spoilage rates than more affluent counties with significant white populations." Further, "Even in counties where the same voting technology was used, blacks were far more likely to have their votes rejected than whites."

As a result of these and other reported voting irregularities, the U.S. Congress enacted the Help America Vote Act (HAVA) of 2002, Public Law 107-252. In the areas related to human factors, usability, and accessibility, the Election Assistance Commission is mandated to submit a report to Congress. Specifically, "…the Commission, in consultation with the Director of the National Institute of Standards and Technology, shall submit a report to Congress which assesses the areas of human factor research, including usability engineering and human-computer and human-machine interaction, which feasibly could be applied to voting products and systems design to ensure the usability and accuracy of voting products and systems, including methods to improve access for individuals with disabilities (including blindness) and individuals with limited proficiency in the English language and to reduce voter error and the number of spoiled ballots in elections." This report was written to address this mandate.

1.3 Brief History of Standards and Testing for Voting Systems[3]

During the 1970s, few states had any guidelines for testing or evaluating voting machines. Stories about voting equipment problems and failures circulated among election officials, triggering concerns about the integrity of the voting process. In 1975, NIST (known then as the National Bureau of Standards, or NBS) prepared a report entitled, *Effective Use of Computing Technology in Vote Tallying* (NBS Special Publication 500-30). The report concluded that one cause of computer-related election problems was the lack of technical skills at the state and local level for developing or implementing complex written standards against which voting system hardware and software could be tested.

This report, along with comments from state and local election officials, led the U.S. Congress to direct the Federal Elections Commission (FEC) to work with NIST to conduct a study of the feasibility of developing national standards for voting systems. Following release of the 1982 report, limited funds were appropriated to begin the multi-year effort. Thirteen meetings and five years later, with the help of about 130 different policy and technical officials, the FEC instituted the 1990 Voluntary Voting System Standards (VSS).

No Federal agency at that point had been assigned responsibility for testing voting equipment against the VSS. The National Association of State Election Directors (NASED) subsequently established a "certification" program through which equipment could be submitted by the vendors to an Independent Testing Authority (ITA) for system qualification. The ITAs are accredited by NASED to determine whether voting products are in compliance with the VSS. The results of the qualification tests can be used by States and local jurisdictions to help them assess system integrity, accuracy, and reliability, as part of their own certification testing.

[3] Thanks to Penelope Bonsall for her help in accurately summarizing the history of the VSS.

The VSS themselves were substantially updated and issued again in 2002, following a three-year development and public review process. This most recent update was accorded favorable review by the General Accounting Office in its preliminary audit (GAO, 2001). This release included functional requirements to improve accessibility by individuals with disabilities. An advisory section was included as guidance to improve user interface and ballot design. There were no specific qualification test criteria developed for this section; hence no formal conformance tests are associated with the guidance.[4]

[4] See Section 2.4 for a more detailed discussion of conformance testing.

2

The purpose of this section is to explain the terminology and concepts of usability, accessibility, standards and conformance testing as used throughout this report.

2.1 Definition of a "System"

The term "system" is used in industry in a number of different ways. Software is often referred to as a system, particularly by those developing software products (e.g., the operating system). A computer is often referred to as a system, though it contains both hardware and software. The hardware, software, and wiring used to interconnect a set of computers are often referred to together as a system (e.g., the networking system). These definitions are problematic in a discussion of usability.

In the usability field, the definition of system encompasses the users and all the elements required to accomplish some goal. A specific system is viewed as one (or more) users, attempting to accomplish some activities towards a goal or set of goals, within a specific environment. The activities include all interaction between the user and other parts of the system (the products, the environment, etc.) as well as activities they might do internally, such as decision-making. Various elements of the environment include: (1) the physical environment (lighting, temperature, and ambient noise), (2) the psychological environment (time or social pressure present in the environment), (3) all of the equipment used, and (4) any other users or support personnel involved. For voting, this means that, from a usability perspective, the voting system is defined by:

- The voters themselves
- The physical environment in which they vote (polling station or home for Internet-based voting)
- The psychological environment associated with voting (e.g., stress induced by long lines at polling stations, social pressures, time pressure associated with personal or other deadlines, etc.)
- The equipment, both hardware and software, used for voting (paper-based voting products, computer-based voting products, etc.)
- The ballot itself
- Quality of support provided (if required by the voter) by poll workers

- Any documentation and training provided (either to the voter or the poll workers and other election administrators)

A change in any of these elements will redefine the system from a usability perspective. Most significantly from the usability perspective, the characteristics of the user population are significant in an evaluation of system usability. These characteristics include:

- Age and gender
- Background (educational, social, and cultural)
- Physical or mental capabilities
- Psychological factors such as current levels of, and susceptibility to changes in, stress, fatigue, mood, and motivation
- Prior experience with the subject matter
- Prior experience with the equipment to be used

These are the "human factors" of the system. Usability is determined by the demands (both physical and psychological) that the other components of the system put on the users and the users' ability to perform under these demands.

2.2 Definitions of Accessibility and Usability

In this report, we have tried to adhere as much as possible to the International Organization for Standardization (ISO) definitions of accessibility and usability. These definitions support development of standards that will lend themselves to suitable test methods for conformance. It is critical to be able to measure accessibility and usability in order to say with authority that a voting product or system has achieved a specified level of accessibility and usability.

2.2.1 Disability

A disability is defined as "a mental or physical impairment which substantially limits one or more of a person's major life activities" by the Americans with Disabilities Act of 1990. This includes, but is not limited to, four major types of impairments: (1) physical impairment such as limited or total loss of use of one or more limbs, limited strength or dexterity, speech impediments, and difficulties in motor control (including tremors), (2) visual impairments ranging from partially to legally blind to total loss of vision as well as other visual deficiencies including color blindness, macular degeneration and tunnel vision, (3) auditory impairments including partial hearing loss in segments of the auditory spectrum or across the entire auditory spectrum, and deafness, and (4) cognitive impairment including learning and reading disabilities, and, under some definitions, users limited in their English proficiency (LEP). There are also common forms of multiple

disabilities (e.g., deaf-blind). Designing a product that could be used unaided by the total range of disabled users (including multiple disabilities) is most likely infeasible, but significant ranges of the populations with disabilities can be accommodated with the proper application of modern technology and good, universal design. This is one of the key areas where computer-based solutions hold significant promise. For example, alternate media are possible such as text-equivalent speech since audio output is considered to be a nearly universal solution for those with visual impairments.[5]

2.2.2 Accessibility

Accessibility is defined as a measurable characteristic: the degree to which a system is available to and usable by individuals with disabilities. The most common disabilities include those associated with vision, hearing, and mobility, but the definition also includes cognitive disabilities. The HAVA also includes accessibility requirements for Native American and Alaska Native citizens and alternative language access for voters with limited English proficiency (LEP).

Accessibility standards are typically intended to specify designs that will maximize the access of the majority of persons with these types of disabilities, but does not necessarily guarantee access for a specific individual's disability or combination of disabilities. An example of this approach to accessibility standards is the set of Section 508 Standards (Section 508 Standards, 2000) developed by the U.S. Access Board for Section 508 of the Rehabilitation Act of 1973, as amended in 1998. The U.S. Access Board is an independent Federal agency devoted to accessible design for people with disabilities. Section 508 is a set of accessibility requirements for Federal electronic and information technology. It applies to all Federal agencies when they procure, develop, use or maintain such technology. Accessibility as defined by the Access Board "is a term that describes products or services that meet the Access Board guidelines (in the case of the ADA) or the standards (in the case of 508). Something that is accessible – i.e., meets the guidelines or standards – is not always usable." The Access Board also recognizes that products are accessible to individuals or groups of individuals. They are never "accessible" as an absolute unless every single person with any type, degree or combination of disabilities would be able to use the products. Products can meet accessibility standards or guidelines. These products are sometimes referred to as "accessible" in this more limited sense. However, these products may still be inaccessible to some people.

The ISO standard TS 16071 defines accessibility as the *usability* [italics added] of a product, service, environment or facility by people with the widest range of capabilities (ISO/TS 16071, 2003). For the purposes of this report, however, we intentionally make a distinction between accessibility and usability since, in general, meeting accessibility standards does not necessarily imply that a system

[5] Reading speed with braille and other tactical displays is significantly slower than even audio output at normal speed. In addition, only a limited number of blind and visually impaired users are proficient in reading braille or other tactile displays.

is usable by a particular individual or even a group of individuals, but only that barriers to access have been removed. This distinction is particularly pertinent to this report and the recommendations for addressing these issues (including the means of testing and certification) are discussed in more detail in Section 6.

2.2.3 Usability and Usability Testing

Usability, for the purposes of this report, is a measure of the effectiveness, efficiency, and satisfaction achieved by a specified set of users performing specified tasks with a given product (ISO 9241-11, 1998). Effectiveness is the accuracy and completeness with which specified users can achieve specified goals in particular environments. Efficiency is defined as the resources expended by the user in relation to the accuracy and completeness of goals achieved. Satisfaction is defined as the subjective comfort and acceptability of the results to its users and other people affected by the results. These definitions have been formulated to provide the means for explicit measurements for usability.

Usability testing is a method by which users of a product are asked to perform certain tasks in an effort to measure the product's usability using the metrics of effectiveness, efficiency, and satisfaction. In practice, usability testing is part of a larger set of approaches for evaluating usability, some of which involve users directly and other which do not. Testing is usually separated into formative (or diagnostic) testing and summative (or empirical) testing. Typically, formative (or diagnostic) testing is conducted as part of a product development process while summative (or empirical) testing is conducted after a product is completed.

2.2.4 Self-Contained, Closed, Accessible Products

For some types of products, it is the responsibility of the user to provide accessibility-related software or an assistive technology device to make the product, sometimes called an “open architecture product”, accessible. In these cases, the designers are responsible for ensuring that their products are compatible with assistive technology. Examples include a sip-and-puff switch used by people who are quadriplegic, or a screen magnifier, screen reader software, braille display, or Opticon[6] used by people who have visual impairments. Voting stations in use at polling places are considered to be “self-contained, closed products” in that they are intended to be used without requiring the user to install specific accessibility-related software or the attachment of an assistive device.[7] This does not preclude the option for users to provide some device such as a mouth stick for pointing or their personal audio headset. It should be noted that some potential voting solutions being proposed such as telephone-based or Internet voting systems would not be designed as self-contained, closed products.

[6] The Opticon is a device that converts visual data into tactile data and can be used to read data from a computer screen. Only a small percentage of blind users use Opticons for this purpose and the reading speed is significantly lower than other forms of alternate output.

[7] This is the term used by the U.S. Access Board. It should not be confused with the concept of an open or closed architecture as used when referring to computer systems.

2.2.5 Accessibility versus Usability

Although the general definition of accessibility includes both availability and usability by people with disabilities, in this report we will treat accessibility as the degree to which a system is *available* to people with disabilities. Access alone does not guarantee usability. The usability of a product by people with disabilities will be considered a subset of the general concept of usability. This view facilitates the development of our recommendations.

2.2.6 Usability in Practice (an Example)

As an example of applying these definitions, consider the typical credit card scanner at the grocery checkout line. Grocery stores, in general, are not very accessible to some disabled populations, so it is not surprising to see that the credit card scanners are not accessible. The scanners are too high to reach and see from a wheel chair, and their combination user interface of push buttons and screen are not readable by people with visual impairments. One could imagine some design guidelines on placement and audio capability that would make the scanner accessible.

Even assuming that customers can access the scanner, there are some obvious usability issues, especially with first- and second-generation scanner designs. The first difficulty is determining how to insert the credit card. Often the diagram next to the slot is horizontal, but the customer must insert the card vertically and must figure out how to match the diagram. To ensure both efficiency and satisfaction, the cashier may take the card and insert it for the customer if the customer tries a few times and still doesn't orient the card properly.

Next there are several options to choose from following instructions on a small screen. The wording and orientation of the instructions on the screen and buttons can be confusing. The dollar total is displayed along with a question of yes or no. But the displayed "yes" and "no" are often located at the top of the screen, quite near to a set of buttons just above the screen for credit or debit. As a result, even with the color coding used on the "yes" and "no" buttons it is not uncommon to see users hesitate or even hit the credit or debit buttons instead of the "yes" and "no" buttons at the bottom, below the screen.

Effectiveness for this product could be measured by different criteria including counting the number of errors (e.g., the number of incorrect insertions of the credit cards or the number of incorrect button presses) or by counting the number of help incidents (e.g., the number of times the cashier intervenes). Efficiency can also be measured several ways including timing the transaction or by counting the number of discrete steps. It can also be measured either only in the "happy path" (e.g., with no errors affecting effectiveness) or as an average over actual use (e.g., including errors affecting effectiveness). Finally, satisfaction can be measured by various means such as through a questionnaire that asks how the customers perceived the process and the outcome (e.g., did they feel frustrated or embarrassed by needing help or by holding up the line). Sufficient

usability testing can be conducted to produce measures of the usability of such a product once suitable criteria are established. The resulting measurements can be used as benchmarks for testing new designs or comparing across products of equivalent capability

2.3 Product Requirements, Usability, and Testing Methods

Product requirements are used to specify in advance what is expected of a product. Requirements may vary greatly in their level of detail and formality: a company may compile a simple list of desired features for its own use, or an authorized committee may compose a lengthy formal standard for use by the general public. In this section, we discuss four independent properties of requirements: their type, their pertinence to human interaction, their level, and their specificity. We define each of these properties and then discuss the implications for usability and for appropriate test methodology.

2.3.1 Type

The two basic types are performance requirements and design requirements (Hemenway, 1980). Performance requirements specify what functions and sub-functions a system is capable of supporting (i.e. the system is analyzed in terms of what operations can be performed), and design requirements specify how the mechanism is designed (physical components and sub components).

Performance requirements can be further subdivided into purely functional requirements and those specifying the degree of performance. As an example, suppose we wish to require that an automobile allow the driver to open the trunk from within the passenger compartment: A design requirement might specify that there be a trunk-release handle at least 4 inches long located no farther than 7 inches to the left of the driver's left knee and that the driver must pull the handle towards the back of the car in order to operate it. A functional (performance) requirement might simply state that there must be a way for the driver to open the trunk without leaving his/her seat. A degree-of-performance requirement might add that the operation must be able to be performed in four seconds or less (on average) once a driver is shown how to operate the mechanism. Note that although this particular example involves human use, the performance/design distinction applies just as well to requirements for autonomous products. See the next section (2.3.2) for more discussion on the human interaction.

Generally speaking, design requirements are appropriate when the purpose of the requirement is interoperability, since this often requires an "exact fit" between system components. While there may be some variation in the concrete implementation of a design requirement (e.g. the color and exact shape of the trunk release handle), the thrust is to constrain the product.

Performance requirements are usually preferable when quality is the goal, since they directly describe the behavior of the system and not the supporting mechanism. Thus, performance requirements allow for innovative solutions, and

they enable comparison of multiple competing designs since they are “technology agnostic”.

Design requirements usually invite direct examination as the most suitable mode of testing. Examination may be as simple as observing the presence of some required part, or may involve very precise measurement and analysis (e.g. analyzing a file to see if it conforms to a format requirement; checking “legal HTML” for a web page is a form of examination).

Performance requirements, on the other hand, are more suited to testing by operation. Since the requirement describes how a system should behave, we operate the system and see whether it works as expected.

2.3.2 Human Interaction

The next distinction is between requirements that have direct and significant effects on interaction between artifacts and human beings, and those concerned with more or less autonomous objects or behavior. In the example above, we saw several different ways in which the ability (of a human) to open the trunk of a car could be mandated by a requirement. Even though these specifications were of different types (design, functional, degree-of-performance), they all had implications for how a human and machine would interact.

Conversely, requirements limiting automobile emissions or describing the format of a DVD do not have direct implications for the way in which humans use automobiles or DVD players.

Interaction requirements can be physical or psychological. Physical interaction requirements are related to such things as the product weight; size; the spacing of knobs, controls, and buttons; the force required to interact with knobs, controls, and buttons; the user’s reach envelope (what parts of the product the user can physically reach); and the user’s field of view (what parts of the product the user can visually “reach”). Psychological interaction requirements include the user’s understanding of the system’s displays, labels, and messages, the user’s understanding of the processes and procedures required to use the product, and the user’s ability to understand the outcome of the interaction (including awareness that goals were met).

Interaction requirements can be formulated in functional terms (e.g. users must be able carry a product) or as design specifications (e.g. the unit’s maximum weight and size). It is presumed that meeting the design specification will ensure the usability of the resultant product. However, design requirements are nearly always based on an “average” or “typical” range of users and do not necessarily apply to all individual users.

Conformance tests for requirements that do not involve human interactions are usually susceptible to automation, since only the intrinsic structure or behavior of some artifact is being tested. When a requirement does involve human interaction, the way in which it is to be tested depends on its type, as defined in Section 2.3.1.

2.3.2.1 Testing Design Requirements for Interactive Products

A test for a design specification (such as the configuration of a trunk release lever) can be done by examination, because even though we expect the mechanism to be used by a human, the requirement has, for better or worse, dictated the design of the "machine" side of the interaction. As long as the mechanism follows the design, it conforms to the requirement, whether or not the design itself is well-suited for the general purpose.

2.3.2.2 Testing Functional Requirements for Interactive Products

A test for a simple "functional capability" specification would normally involve operation by an expert, presumably by following the instructions for the product in question. The expert would check for the presence and general workability of a mechanism supporting the capability. For example, if the examiner can indeed open the trunk from the driver's seat, the test is passed. As with design specifications, passing the test is no guarantee of real usability. A product could actually include a mechanism to accomplish some task, and yet have poor usability if the mechanism is hard to understand or difficult to operate. See Section 3.2.4 for an extended example of how usability can be affected by various designs for the same function on a voting product.

2.3.2.3 Testing Performance Requirements for Interactive Products

Finally, a direct test of a degree-of-performance specification would normally involve some use of human subjects. For example, if the requirement specifies how long it takes a certain class of drivers to open the trunk after being instructed, the only real way to tell if a given automobile conforms is to have some users try it out. Of course, much hinges on the exact wording of the requirement (possible metrics include: the time users are expected to take to perform the activities on the system, the throughput of the system such as the number of forms that are processed in a given amount of time, the number of acceptable errors a user can perform in typical interaction) and how precise we wish the test to be. A casual test might substitute expert judgment for actual experimentation. Nonetheless, if the requirement mandates a certain degree of success by a designated class of users, a usability test of some sort is the most direct and reliable way of measuring conformance.

A special case of a performance specification would be to mandate a certain degree of user satisfaction with the use of the product. Within the usability and human factors engineering community, user satisfaction is often considered a requirement for a product, but it is rarely stated specifically. However, this trend is changing. Because satisfaction is one dimension of the system's usability that is

purely subjective, the requirements for satisfaction are difficult, but not impossible, to include in product testing. Testing might involve questionnaires or interviews with users to determine their subjective reaction to the use of the product. Note that a number of validated subjective satisfaction questionnaires exist, such as SUMI (http://sumi.ucc.ie/), SUS (http://www.cee.hw.ac.uk/~ph/sus.html), and QUIS (http://www.cs.umd.edu/hcil/quis/).

2.3.3 Levels

The various specifications within a requirement can address the entire system in question (high-level) or major components or functions of the system (mid-level) or small components and functions (low-level). These are not precise distinctions. For example, the acceleration of an automobile is a high-level function; the ability to open the trunk from the engine compartment is a mid-to-low level function. Low-level requirements are often lengthy and embody a detailed description of the object in question. Design requirements tend to be low-level although performance requirements can also be low-level. Higher-level requirements embody more abstract requirements; the basic functions or mechanisms are specified without dictating the details of how this is accomplished. Note, in particular, that high-level performance requirements directly address the broad "bottom-line" requirements of a system, without constraining the means by which these are achieved.

The testing implications are straightforward: numerous low-level specifications typically require numerous tests (although each test is likely to be small and simple). Higher-level specifications might require somewhat more complex and holistic tests, but there are likely to be fewer of them.

2.3.4 Specificity

Requirements can be couched in terms that range from the very specific to the general. For example, at one extreme, a requirement might simply specify in general that an automobile "provide good visibility" for the driver; at the other, the requirement might specifically mandate the precise angular height and width of an unobstructed view for any driver between a minimum and maximum height.

Specific requirements are not the only mechanism by which quality can be encouraged. In cases where it is impossible to state precise requirements, general specifications can provide useful guidance. Indeed, some usability "requirements" (e.g., use legible fonts) are in fact simply checklists of general design considerations to be taken into account by developers.

Specificity is important when it is necessary to test objectively whether a given system conforms to a requirement. General specifications lead to tests that are either subjective (e.g. an expert uses his/her judgment to decide whether the visibility is "good") or somewhat arbitrary (the test procedure, rather than the requirement, adopts a more precise definition of what constitutes good visibility).

Specific requirements support test procedures that are both more objective and more directly justified by the text of the requirement.

2.4 Standards and Conformance Testing

Standards are a ubiquitous part of the "invisible infrastructure" that helps to assure that functions such as commerce, transportation, and communication take place smoothly and integrate appropriately. Standards can be formulated and applied in various ways. The following is a brief overview of the basic concepts of standardization.

2.4.1 Terminology of Standards

A standard usually has one of three basic purposes:

- To provide a metric for the accurate measurement of some property, such as the unit for mass (maintained by NIST), namely the kilogram. The standard is used for comparison so that all measures of a given unit are equivalent.

- To assure the interoperation of manufactured components of a system, such as the format for a compact disk to be read by a CD player.

- To establish a level of quality to be met by a product, such as emission standards for automobiles.

In this report, our main concerns are the standards of quality for usability and accessibility in voting and interoperability standards for some accessibility issues. Standards for these latter two purposes are really just examples of detailed, officially formulated product requirements, as described above in Section 2.3. (Standards for metrics are a different case and not of concern in this report.) As such, all the remarks above concerning the properties of requirements (type, level, etc.) and the implications for testing apply directly to interoperability standards and quality standards.

2.4.2 Pragmatic Issues for the Application of Standards

Once a standard is approved, three other issues emerge:

- How to resolve disputes about the meaning of the standard: the **interpretation** problem,

- How to tell whether a given entity conforms to its requirements: the **testing** problem, and

- What happens if a product doesn't conform: the **enforcement** problem.

2.4.2.1 Interpretation

As with any written requirement, standards can become the subject of dispute, especially when the conformance of a product depends on the precise interpretation of the standard's wording. This is a good argument for writing clear and specific requirements in the first place. Nonetheless, disputes do arise, and some authoritative body must decide the issue. This "judicial" function has been carried out in various ways. Usually, the body that developed the standard also takes the responsibility for providing interpretations, but sometimes a third party will be given this task.

2.4.2.2 Conformance Testing

If a standard is to be something more than a mere document, there must be some procedure for applying the standard to the entities within its scope. This procedure must be designed and written (test development) and then enacted (test operation). The development of a good test suite can often involve more effort than the formulation of the standard itself. As we have seen, the test methodology depends strongly on all four properties (type, human interaction, level, specificity) of the standard.

Note that there are many types of testing that do not deal directly with conformance – examples include:

- Exploratory testing in the early design stage of development (usually called formative testing in the usability field),
- Debugging (diagnostic testing for defects), and
- Comparative testing of competing products.

In particular, note that although conformance tests may often have some diagnostic value, **their main purpose is to detect aspects of the system that do and do not meet the requirements of the standard**, not to find the cause of the failure.

Finally, there is the issue of test operation. Conformance test suites are sometimes executed by the vendor (self-testing), or by a potential purchaser. It has also become common practice for a third party, such as an accredited laboratory to perform the testing. As mentioned earlier, such third-party testers are referred to as Independent Testing Authorities (ITAs).

2.4.2.3 Enforcement

There are two points of decision for enforcement, the first internal to the standard, the second external.

In the first case, the standard itself may make distinctions between its binding specifications (often denoted by saying that something “shall” be the case) and non-binding specifications (denoted by “should”).

Second, the standard as a whole may be enforced in several ways:

- It may be enforced directly by the Government,
- It may be enforced by potential buyers who refuse to purchase non-conforming products,
- There may be a “labeling” policy, such that a product cannot claim to be of a certain type, unless it meets the standard, or
- It may be completely voluntary.

The enforcement approach taken is a policy issue, and normally has no direct technical implications.

3 Usability and Accessibility Requirements of Voting

As one would expect, the various kinds of usability-related requirements are well-represented for voting products, though most are functional- and performance-based. For example, there is a functional requirement for the voter to have the ability to cast a single vote in a winner-take-all election or to cast multiple votes in a multi-member election. There is a functional requirement to allow voters to modify their votes before casting them. Functional requirements also provide constraints on the interactions, designed to protect the voter from inadvertent errors such as the provisions to prevent overvotes and to notify voters of undervotes. There are also general functional requirements such as the ability for voters with disabilities to interact with the product.

Interaction requirements can also be identified for voting products some of which exist in current or draft standards. These include specification of the typical reach envelope, minimum font size, and other specific design details. As with all design requirements, there is some question as to the effect these requirements actually have on the usability of the product.

User performance requirements for voting products also exist. These are generally not enforced and appear to be provided only as guidelines. In one state, for example, there is a requirement that the act of voting by an individual voter take no more than five minutes. However, there does not appear to be any currently defined requirements related to user error rates, though there appears to be an implied requirement related to DRE systems that sets the error rate for overvoting to zero. However, there appears to be no standard for the number of anticipated user errors, or the number of calls for assistance that would be considered acceptable when dealing with a large user population such as is the case with voting. Such errors might include the number of times a voter inadvertently attempts to overvote, unintentionally undervotes a ballot, or is unsure of the next step in a process, whether these conditions are corrected or not before the vote is cast.

User satisfaction requirements do not appear to be defined for voting products though they have been the subject of many articles on voting.

The question that remains to be answered is whether or not existing standards are necessary and/or sufficient to ensure a high degree of usability for voting products. Finally, it is important to note that there is considerable variation in the implementation and design of voting products, which makes it a challenging task to create standards that are testable, span this range of design, and ensure some level of good usability and accessibility.

3.1 Implementation Examples of Functional Requirements for Voting

3.1.1 Implementation Variations

As described above, the specific implementation developed by a vendor defines the interaction characteristics of the product. And, it is these interaction characteristics that determine the usability and accessibility of a product. Multiple implementations are nearly always possible for a set of functional requirements and the implementation variations are often based on variations in the medium or technology used. In the case of voting products, examples are the selection of paper versus electronic, touch screen versus selection wheel, and QWERTY versus alphabetic keyboards for write-ins. The choice of technology is up to the designer, though much of it may be limited by technological capabilities, cost, or some other factor. The choice of the interface design, including many elements of the presentation layout and screen flow is also up to the designer, though it is constrained somewhat by standards, conventions, and industry best practices. Assuming all functional requirements are supported, variations in the specific implementations will cause different interaction challenges.

Since multiple designs may be in use across the country, across the state, or even across a district, the usability of the products will vary. Further, within the U.S., vendors are free to create unique voting products. This is due, in part, to our culture of both voluntary standards and free competition. Contrast this with the approach taken by the Brazilian government, which contracted with two research companies to design a single product for voting (Caltech–MIT, 2001). Separate contracts were made with a number of companies to manufacture the product to the design specifications. In this situation, provided the product is both accessible and usable, all voters will experience the same system and therefore the same level of usability and accessibility will be seen across the entire country.

Although this single design approach is a possibility for the U.S, the variations in State requirements and the nature of the relationship between the Federal government and State governments make it a highly unlikely solution. Nevertheless, it is still necessary to ensure that all designs from all vendors achieve a minimum level of usability and accessibility. The current VSS approach assumes that appropriate standards can be put in place to ensure the usability and accessibility of voting products. However, design standards can ensure a specific level of usability and accessibility only if they completely specify the interface design. This can restrict both the incorporation of new advances in technology as well as creativity on the part of the designers to develop novel solutions. Alternatively, the usability and accessibility of each product can be independently determined and compared to a fixed standard for these aspects of the product design. It is for this reason that this report focuses on performance-based standards for both usability and accessibility and minimizes the dependence on design standards.

3.1.2 Example of Voting Product Design Variations

One functional requirement from the HAVA requires that voting systems allow the voters to change their votes in any of the contests before casting their ballots. In a review of the existing voting products at the 2003 International Association of Clerks, Recorders, Election Officials and Treasurers (IACREOT) Trade Show in Denver, we observed at least three different implementations of this requirement. All three were touch screen-based, DRE products and all three technically met the current VSS standards. Yet significant variations in the designs and implementations existed. Note that the variations themselves do not necessarily indicate a problem with existing standards or the DREs themselves, but do indicate the importance of measuring usability. Further, although we have focused on DREs, similar issues can be identified for all systems. (For example, paper ballots do not prevent overvoting and lever machines, in some cases, had labels too high for small individuals.)

In this section, we will describe these designs in terms of their interaction characteristics[8] in a winner-takes-all type election and discuss the resulting usability issues.

3.1.2.1 Product A – No Change Feedback

In one design, the voter selects a name from a list of candidates by pressing on the name on the screen through the system's touch screen interface. If he touches a different name within the same contest, the first choice is changed to the new selection. There are no messages (auditory or visual) associated with this change action beyond the highlight of the new choice.

3.1.2.2 Product B – Yes/No on Change

In a second design, the voter selects a name using the same touch screen approach but, if the voter presses on the second name, a message is displayed asking her if she is attempting to change her vote. The voter can select "yes", in which case the system removes the message from the screen, removes the mark from the first candidate, and marks the second candidate as the selected choice. If the voter selects "no", the system removes the message from the screen, but no other action is taken.

3.1.2.3 Product C – Deselect/Select to Change

In a third design, the voter makes a selection for the first candidate as he would do with the other two designs but, if he presses on the second name, nothing happens. There is no change in his vote and no message displayed. In this design, the voter must reselect the first candidate again to remove the selection mark before he can select a new name.

[8] Note that in describing the voter interaction in these examples we have chosen to alternate the use of the pronouns "he" and "she" to give some indication of the diversity of voters and paint a more vivid picture of the user interaction.

3.1.2.4 Interpretation from a Usability Perspective

These three designs represent different approaches to satisfying the same functional requirement using basically equivalent technology (e.g., a touch screen-based, DRE interface). Other technologies are possible as well. For example, another voting product also reviewed at the trade show had the same interaction characteristics as Product C but used a non-touch screen interface for selection. In this case, the navigation and selection was accomplished through a rotating wheel and a selection button and no other means of navigation and selection were provided.

Product A appears to be the simplest to use since it contains the fewest number of steps. User action is responded to directly by the system. However, this design includes the possibility of the voter inadvertently changing his vote and not detecting this. If, while moving his hand across the screen, the voter accidentally touches the name of an alternate candidate that candidate will be selected. If the voter fails to notice this change, and continues the voting process, he won't necessarily notice this error during the review. If he does notice this error, he must return to the selection and change his vote. Even if the voter is able to perform this pass without difficulty (he is able to determine how to return to the contest and make the correction), it will increase the time on task for this voter. The potential also exists for the voter to fail to notice this change, even during the review process, and cast his votes with the inadvertent error.

Product B has a specific feature apparently designed to prevent this very error. The voter must specifically acknowledge the change in vote before it takes effect.

Product C appears also to prevent inadvertent selection of an alternate vote, but does so in a fashion that requires the voter to determine why the system failed to respond. He must determine that the vote has to be removed before the intended vote can be cast. There is some question as to whether or not the voter would realize this by himself. Some voters might have no difficulty in making this determination. Others might ask for help to understand how to make this change. Some voters might perceive (incorrectly) that the system does not allow them to make the change once a candidate selection is shown on the screen.

All three designs support the functional requirement but provide separate usability challenges for the voters in terms of what they must physically do and mentally understand. As a result, the error rates for each of these designs (in terms of voter confusion, calls for help, error rate, time on task, or voter acceptance) would likely vary, though all three errors rates may very well be within acceptable limits. The actual error rate of these designs cannot be determined without adequate testing with actual voters and a determination of what are "acceptable" limits.

3.2 Potential Usability Problems in Voting Products

In this section, we discuss the types of usability issues associated with voting products. We assume here that there are no accessibility issues and these discussions apply to all voters.

There are a number of factors that determine the nature and frequency of usability problems encountered with any product. Users must be able to (1) deduce the interaction required (or be trained and able to accurately recall the interaction) and (2) be physically and mentally able to perform the interaction. If users cannot achieve (1) or (2), they will not be able to use the system. However, since we are talking about human involvement, perfect usability is rarely if ever realized. Instead, usability varies between perfect success (a perfect match between designer expectation and user interaction and accomplishment of the goal within an allotted or acceptable period of time) and total failure (the inability to reach the goal or to reach the goal accurately within an allotted or acceptable period of time).

Hence, there are three classes of usability problems:

- Usability problems prior to success
- Usability problems prior to partial failure
- Usability problems leading to total failure

3.2.1 Usability Problems Prior to Success

Usability problems can affect voters during the process but not affect their ability to accomplish the goal of casting a valid vote as intended. These problems will result in changes in the overall task time. Problems might also show up as voters try to perform inappropriate actions such as attempting to move or scroll beyond the limits of a page, accidentally changing votes, attempting to overvote a contest, or unintentionally undervoting an election. Provided the voter corrects each of these errors before completing the goal of casting a ballot, the results can still be considered successful, even though it took additional time or used more cognitive or physical effort than it would have if the voter's actions had taken place with an alternate design. (Note that if large numbers of voters take extra time, this could result in longer waiting lines at the polling place discouraging some voters from staying to vote.) In addition, the problems experienced in attempting the task of voting may be severe enough to require assistance (either on line or live). Finally, they can result in changes in the user's level of satisfaction. [9] When dealing with a large number of voters, some will inevitably struggle mentally or physically with the interface before determining the correct interaction required but will ultimately

[9] It is common for users to blame themselves for their inability to accomplish a task with a given system, even though difficulties experienced may be common across a range of users and the result of correctable usability problems.

succeed. Others will make and correct one or more process errors before succeeding.

In a voting product, these usability problems might manifest themselves by increases in the time it takes one user to vote. These changes can be detected only by measuring the actual time required to vote or by directly observing voter behaviors (e.g., physical hesitation while voting). From an individual perspective, the task is still completed so the issues might not be serious enough to address, except where problems affect the user's subjective ratings including confidence in the final result. Although individual performance might not be sufficiently affected to warrant concern over the design, additional delays in lines and added frustration on the part of those waiting can be severe enough to affect the overall system performance (i.e., the collection of all voters in a given polling location). Voters waiting longer in line may perform worse than those that have to wait less (potentially leading to more frequent or more severe usability problems) or might even leave before voting.

3.2.2 Usability Problems Leading to Partial Failure

More critical usability problems can result in the user being able to accomplish only some of the tasks associated with a goal or exceeding acceptable time limits. These usability problems might also be manifested in changes in individual user behavior, time on task, and user satisfaction, but they would also be show up as changes in the quality of the final product (e.g., the ballots may show more undervotes, null voters within specific elections, or rolloff voting behaviors). Since the tasks that are accomplished are correct, the final result of these problems would be classified as usability problems prior to partial failure. However, it should be noted that the presence of undervotes, null votes within specific elections, or rolloff voting behaviors cannot be assumed to be the results of usability problems. The voter may have intentionally decided to perform these actions.

More disturbing than the presence of problems leading to partial failure, is the fact that voters might not even be aware of the existence of the problems. For example, a voter may unintentionally cast a ballot that shows signs of rolloff voting behavior believing that they voted all levels of an election when, in fact, they did not. Note also that usability problems may result in added pressure to complete the ballot, thus resulting in a conscious decision not to vote in some races even though this was not the original intention (a usability problem leading to partial failure). This would also be classified as a usability problem leading to partial failure, but the voter would not necessarily view this as a failure. (This is one reason why exit polling and post test surveys can lead to false impressions about the nature and extent of usability problems in a product.)

3.2.3 Usability Problems Leading to Total Failure

A usability problem that resulted in the total inability to perform the tasks or to perform the task within an allotted or acceptable period of time would be considered a usability problem leading to total failure. In a voting product, one presumes this is a rare case since total failure appears to be a detectable event and the voter would likely be assisted by a poll worker to complete the voting process. However, such usability problems may not actually be that rare, just rarely observed, for they can be manifested by a voter prematurely casting a ballot (which was reported in Maryland) or the voter leaving the voting booth without casting the ballot (which was reported in New York). Voters can also become so frustrated that they abandon the voting process without completing a ballot.

3.2.4 Examples of Potential Usability Problems

Using the examples of the three touch screen-based, DRE Products A, B, and C described earlier, we can look at the variations in the designs and speculate about the usability problems (which represent a potential for error), and even predict relative rates (though it is not possible to predict actual rates without a controlled study of actual voters). We assume for the purpose of these illustrations that the voters are physically able to interact with a touch screen. (In general, this should not be the only way that people with disabilities can interact with the product.)

This analysis is based on the physical and psychological interaction required. However, other factors can also affect the nature and probability of error. Variations in the visual display or spacing could also change how errors occur for any of these designs. The sensitivity or internal technology of the different touch screen devices could also change the error profile. Without actual usability testing, we cannot know if any of these designs or potential errors described would cause usability problems prior to success, usability problems leading to partial failure or a false sense of success, or even usability problems leading to failure.

3.2.4.1 Example 1 – Changing a Vote

Using Product A from Section 3.1.2.1 described above, there is a potential for the voter to inadvertently touch the screen and change her vote while moving her hand across the screen. If she detected and corrected the mistake, this would represent a “usability problem prior to success.”

There is also a possibility that this event would not be detected by the voter at the time it occurs and therefore, the error would not be corrected at that point.[10] If the event is not detected at the time it occurred, then she would have another opportunity to detect the event during the ballot review (as mandated by HAVA) before casting her vote. However, there is still a possibility that she might not detect the event and cast her ballot with the unintended error. This illustrates a usability problem leading to a false sense of success.

Product B has an additional design feature that appears to specifically preclude the inadvertent selection error and thus would likely have fewer incidents of this error going undetected (since the inadvertent contact with the screen results in a message). However, it introduces something new that the voter must understand and interact with correctly. There is the possibility that he might select "yes" instead of "no" or vice versa, which could be influenced by the arrangement of the buttons, their prior experience with similar messages on computer systems, or the wording of the message.[11,12] Further, such an error message should be "modal", that is, it does not allow the voter to interact with any other part of the system until he completes the interaction with this message. If the error message is not "modal", there is a possibility that the message might accidentally become hidden from view and leave the system in an indeterminate state without actually casting a vote. It is unclear in this specific case if the product would allow the user to cast a ballot with a non-modal message open. If it did, then this illustrates a usability problem leading to total failure.

Product C, where the voter must reselect the first candidate's name to remove the selection mark before selecting a new name, represents a design that has a lower or even zero probability that an inadvertent vote change will occur since it would require inadvertently touching the screen twice – one to remove the existing vote and once to inadvertently select the new vote. However, this same design must support the voter's attempt to change her vote. This design appears to be the most difficult for voters to understand since it lacks any feedback or guidance telling a voter how to change her vote. Thus, if she selects a new candidate without de-selecting the first candidate's name (assuming that this is the correct action required), there is no feedback during the event to alert her to a problem. Rather than assuming her action was inappropriate to accomplish her vote change, she might assume that this design does not allow her to make a new selection once one is made. This would make it a usability problem resulting in partial failure. Even if the voters assume the system *should* allow this type of

[10] There is also the probability that the change is detected, but the user does not know why and assumes it to be a system error. In this case, users may correct the error but be suspect of the ability of the system to accurately capture their vote or they might assume that the presumed error needs to be reported to a poll worker. Alternatively, they might be startled by the error and lose confidence in their ability to operate the system.

[11] Failure to adhere to standard button arrangement for Yes/No messages is one of a number of common errors in design that results in inadvertent activation of the incorrect choice.

[12] One blatant example of this type of error is an application message from a commercial software package that reads "OK to not save changes?"

change (or were told that it did), they might struggle with this design or seek help from a poll worker.

3.2.4.2 Example 2 – Voting a Multi-Seat Contest

All three of the example products support the two interaction requirements that the voter be able to select from multiple names within a multi-seat contest and change his vote before casting his ballot. In all three cases, he selects candidates using the same touch screen selection method that is used for selecting a candidate in a single-seat contest. The visual display of each of these designs is consistent across the three products as well as between the single-seat and multi-seat contest. An un-selected candidate is represented as a closed box and a selected candidate is represented as a closed box with a check mark.

In other words, two different types of contests are represented by the same visual design. Readers familiar with computer programs with graphical user interfaces will note that these different types of behaviors (single selection from a group and multiple selections from a group) are generally represented by *two different* visual elements – a round circle (called a radio button) for single selections from a group and a square for multiple selections from a group (called a check box). These different visual designs are intended to aid the user in visually identifying the capabilities (and their inherent interaction requirements) associated with the type of each group. The design of the example products can appear to voters familiar with computers as internally inconsistent (or even as "coded wrong") and thus might represent a usability problem prior to success. There is also a chance that a voter could mistake the box shape as a computer type check box and attempt to overvote a single-seat election. With the new DRE products this would result in a usability problem prior to success since nearly all of them preclude overvoting. Finally, some users may mistake a multi-seat contest as a single-seat contest and inadvertently undervote the contest – a usability problem leading to partial failure.[13]

3.3 Potential Accessibility Problems in Voting Products

Since we have elected to restrict our definition of accessibility to access to, but not usability of, the product and cover usability by people with disabilities as a subcategory of usability, this section primarily discusses barriers to the accessibility of the product, with only a limited discussion of usability issues.

Accessibility represents a wide range of issues and design challenges. Not only must access be provided to people with many types of disabilities, but also access must be provided for U.S. citizens who are not proficient in English and who have different cultural backgrounds (including Native Americans). For a voting system to be accessible one must first remove barriers to access. Then

[13] Alternate visual designs for single- and multi-seat contests could be used to reduce the probability of inadvertent undervoting by providing feedback in a multi-seat contest of the current number of selected candidates and the maximum number of candidates allowed.

interaction requirements can be addressed as part of a usability analysis to ensure that the system is actually usable by these diverse populations.

To satisfy the goal of accessibility, barriers to access by people with disabilities and language difficulties must be removed or an alternate means of access provided. These barriers are often represented by physical barriers such as the inability to enter a building, to reach controls or read displays from a seated position, to interact with controls that require visual feedback (e.g., touch screens) or to use a mouse or a touch screen due to lack of fine motor control. Difficulty in communication can also be a barrier to access. Products that provide information via audible feedback exclusively may be difficult or impossible to use by persons who are deaf or have hearing loss. The issue is present not only in the primary display but also in the feedback used to indicate progress or selection. Many products provide auditory feedback for the user to indicate the end of a page or the last page in a multi-page form, which is fine if the feedback is redundant and also available as visual information. Touch screen products often use auditory feedback to aid the user in knowing that a selection has been made (though this is nearly always redundant with visual feedback). Visual displays cannot be accessed by many users who have visual impairments. Again, a touch screen product, even if not used for data display, relies heavily on visual feedback for proper operation. Once the barriers to access are removed by adding redundancy, a second condition must be satisfied – the product must be usable by these populations.

There is some interaction between usability and accessibility, since the means of providing access represents interaction challenges for the user. Some environmental factors may have changed from those used by non-disabled users. The product may have a different keyboard or entry device for disabled users. In the case of touch screen-based, DRE products for example, an alternate set of keys typically are provided for movement and selection. There might also be differences in the medium used for data display and feedback (e.g., audio instead of video, text instead of audio, alternate entry device instead of a touch screen). Interestingly, specific accessibility features, if used by non-disabled users, may reduce some usability problems. The usability problems noted in the DRE interface design that allowed users to change votes with only a visual indication of the event risks inadvertent activation. However, a user who is blind, using an audio interface, is provided with the name of the new selection even if the selection was inadvertent. This would increase the probability of detection of the event. However, accessible interfaces are often provided as an alternative and are not integrated. Sighted users might not have access to the audio when they are using the touch screen interface. (Note that voters with vision or cognitive problems can benefit from the audio together with a touch screen to confirm that they are reading and interpreting the screens correctly.) In addition, there are special issues of usability for disabled users. The design of the ballot, the length of the ballot, the number of candidates, and the number of races might be

obvious to a sighted user, but not to a voter who is blind or visually impaired unless a feature is included that provides this information.

Though access may be provided, additional requirements for product usability by people with disabilities exist. For both touch screen and non-touch screen-based DRE products reviewed for this report, audio is the primary alternate medium provided for users who are blind or visually impaired. There are many good reasons for this decision on the part of vendors, but problems remain. Audio may be provided as recorded speech or synthetic speech, each with its own benefits and disadvantages. Audio feedback takes longer than reading unless the user can and is able to understand audio playback at high speed. In any case, audio data is transient, so users who are blind or visually impaired rely on short term memory to a larger extent and for more data then non-disabled users. Browsing an audio display is significantly harder and more time consuming than browsing visual displays. Some voters who are deaf might take a longer time to vote. Deaf individuals, particularly those who are congenitally deaf, read at a lower reading level than non-disabled users.[14] Data entry via an alternate input device may be more difficult, take more steps, or have other differences than the primary input.

At a minimum, a fully accessible user interface is anticipated to have an average longer time on task for a person interacting with an audio-based interface than the time on task using a visual display. For standardized tests, it is presumed that there is a 50% increase in task time. However, this estimate is based on completing a standardized test, at a desk, using a familiar alternative interface. Personal correspondences by the authors of this report with individuals with visual disabilities place the estimate on the order of 3 to 4 times longer for some users who are blind. Furthermore, the nature and frequency of usability problems encountered are almost certain to be different.

[14] This is the often the case for congenitally deaf users (those deaf from birth) since reading ability is learned to a large extent as an auditory process.

4 Current Usability and Accessibility Related Standards

Generic standards exist for usability and accessibility that are available from sources such as standards and professional organizations, as well as military and corporate institutions. In addition, some portion of the existing VSS and proposed IEEE standards for voting systems address some aspects of usability and accessibility. This section reviews these sources.

4.1 Current (and Proposed) Voting Systems Standards related to Human Factors, Usability, and Accessibility

Only recently, in the wake of problems revealed in the 2000 elections, has significant attention turned towards the issues of human factors, usability, and accessibility. There are a number of references to usability and accessibility in the existing VSS and proposed IEEE standards for voting systems. A brief overview of the current (as of October 2003) standards environment follows. The information presented has been gleaned from the following sources:

- Help America Vote Act (HAVA)
- Federal Election Commission's (FEC) Voting System Standards (VSS)
- Institute of Electrical and Electronic Engineers (IEEE).

4.1.1 Requirements of HAVA

Title III of HAVA directly imposes certain requirements by 2006 on voting systems used for elections for Federal offices. Those related to usability include:

- Voter verification of the ballot, in private, before final submission
- Voter opportunity to correct a ballot in private before final submission
- Notification of a potential overvote before casting and allowance for correction in private
- Accessibility for voters with disabilities
- Availability of alternative languages for data presentation
- Public availability of certain voting information, including a sample ballot, and instructions on how to cast a vote

These represent both high-level and mid-level functional requirements.

4.1.2 Current FEC Process: the VSS

The voluntary Voting System Standards (VSS) issued by the Federal Election Commission (FEC) have been in effect since 1990 and were last updated on April 30, 2002. According to HAVA, they continue to be the official Federal standards for voting until superseded by guidelines issued by the newly created Election Assistance Commission (EAC). The VSS is a lengthy document comprising an overview and two volumes, the first containing the standards themselves and the second covering test methods. The current voting system standards contain specifications that are functional, for the most part, but also include some design specifications.

Significantly, the VSS define a voting system as the *devices* that allow users to vote. Voters are not considered part of the system. This is in contrast to the definition provided in this report. There is a consequent emphasis on mechanical and electronic performance of the device. Usability is presently covered only in an advisory Appendix, although there are plans to add it as an official specification. Also, no coverage is included for mail-in or absentee balloting or for Internet voting. Both of these areas may present significant issues in the area of security and the potential for fraud or misuse, but the usability aspects of these areas could be addressed independently. Conversely, there is considerable attention given to telecommunications, again demonstrating the present emphasis on the technical aspects of voting.

In general, the VSS document does not clearly define human factors, usability, accessibility, or the associated conformance testing that should be applied to these areas.

In the following sections, we describe the VSS sections on Human Factors, Usability, and Accessibility in more detail.

4.1.2.1 VSS: Human Factors and Usability

The VSS notes that human factors issues are covered mainly in Appendix C to Volume I. ITAs are given "wide latitude to *develop* [emphasis added] and perform appropriate tests". Human error rates are acknowledged, but are mentioned only in relationship to system error rates as follows:

> "…the term 'error rate' applies to errors introduced by the system and not by a voter's action, such as the failure to mark a ballot in accordance with instructions. ... Further research on human interface and usability issues is needed to enable the development of Standards for error rates that account for human error" (page 5).

In the following subsections, we identify features of the VSS that are pertinent to our discussion of how to develop and test usability and accessibility standards. Note that this discussion is by no means a thorough analysis of the VSS and is

not intended to diminish the tremendous and valuable efforts of the FEC and NASED over the past 25 years to develop these standards.

4.1.2.2 Pertinent Features of VSS: Volume I, Performance Standards

In Volume I of the VSS a broad range of levels and specificity is represented in the standards (see 2.3 of this report). Some specifications are quite precise; others very general. Some specifications are low-level; others are very high-level performance and functional requirements. Some of the standards are technology-specific, and others are generic. Generally speaking, the standards have a "bottom-up" empirical feel to them--as if they were composed in response to various technical issues and problems as they arose. Some sections read more like articles on good practice than true standards in a technical sense. As we will see below, there are some ISO standards that are also essentially advice and checklists.

VSS Section 1.1 emphasizes a non-process oriented approach taken within the standards:

> "For the most part, the Standards address what a voting system should reliably do, not how system components should be configured to meet these requirements. It is not the intent of the Standards to impede the design and development of new, innovative equipment by vendors."

This is broadly true since most of the specifications are functional requirements in the form, "The system must be able to do X".

VSS Section 1.5.1 contains the definition of a voting system:

> "A voting system is a combination of mechanical, electromechanical, or electronic equipment. It includes the software required to program, control, and support the equipment that is used to define ballots; to cast and count votes; to report and/or display election results; and to maintain and produce all audit trail information. A voting system may also include the transmission of results over telecommunication networks."

VSS Section 1.6: clarifies the types of testing used for voting systems:

> "...voting systems are subject to the following three testing phases prior to being purchased or leased"
>
> - Qualification tests [performed by the ITA],
> - State certification tests, and
> - State and/or local acceptance tests."

VSS Section 2 is central to understanding the Voting System Standards. Here the general functionality expected of a voting system is defined. Even though most of it does not directly address usability, it does convey the general approach of the VSS to standardization. VSS Section 2.1 commits to functional-style standards: "This section sets out precisely what it is that a voting system is required to do." Indeed, most of the requirements are functional, but there are a few low-level design specifications as well.

Accessibility is covered in VSS Section 2.2.7. Here some very specific design standards are given. For example: "Where any operable control is 10 inches or less behind the reference plane, [the system shall] have a height that is between 15 inches and 54 inches above the floor." This section has a wide variety of types of standards, ranging from broad functional statements to narrow and technology-specific design requirements, many of which can be subject to broad interpretation. For example, in the VSS:

- Section 2.3.1 Ballot Preparation states: "Ensuring that vote response fields, selection buttons, or switches properly align with the specific candidate names and/or issues printed on the ballot display".

- Section 2.4.3.1 notes that all systems shall: "...provide text that is at least 3 millimeters high and provide the capability to adjust or magnify the text to an apparent size of 6.3 millimeters."

- Section 2.4.3.2.1 states: "All paper-based systems shall… ...allow the voter to easily identify the voting field that is associated with each candidate or ballot measure response."

- Section 2.4.3.2.2 states that all "paper-based precinct count systems shall… ...provide feedback to the voter that identifies specific contests or ballot issues for which an overvote or undervote is detected;"

- Section 2.4.3.3 (DRE System Standards) states the system shall "enable the voter to easily identify the selection button or switch, or the active area of the ballot display that is associated with each candidate or ballot measure response;" This section also requires feedback for over- and undervoting and the ability to delete or change choices. And the system must "…Provide sufficient computational performance to provide responses back to each voter entry in no more than three seconds;"

Beyond VSS Section 2, a few more usability related issues are mentioned in various places:

Section 3.2.4.1 states that "all systems" shall provide "...privacy for the voter, and be designed in such a way as to prevent observation of the ballot by any person other than the voter".

Section 3.2.4.2.2 states that: "punching devices" shall "...facilitate the clear and accurate recording of each vote intended by the voter".

In contrast to the surrounding material, there is a short subsection, VSS Section 3.4.9, on Human Engineering – Controls and Display that contains explicit functional and design requirements for usability. It begins:

> "All voting systems and components shall be designed and constructed so as to *simplify and facilitate the functions required*, and to *eliminate the likelihood of erroneous stimuli* and responses on the part of the voter or operator." [Emphasis added.]

This section goes on to state that: "Appendix C provides additional *advisory guidance* on the application of human engineering principles to the interface between the voter and the voting system."

VSS Section 9 on Qualification Testing (a term equivalent to conformance testing as it is described in Section 2.4 of this report) describes a general approach; however, no criteria or procedures for usability and accessibility testing are specified:

> "Qualification testing encompasses the examination of software; tests of hardware under conditions simulating the intended storage, operation, transportation, and maintenance environments; the inspection and evaluation of system documentation; and operational tests to validate system performance and function under normal and abnormal conditions. The testing also evaluates the completeness of the vendor's developmental test program, including the sufficiency of vendor tests conducted to demonstrate compliance with stated system design and performance specifications, and the vendor's documented quality assurance and configuration management practices. The tests address individual system components or elements, as well as the integrated system as a whole…
>
> *Qualification testing is distinct from all other forms of testing*, [emphasis added] including developmental testing by the vendor, certification testing by a state election organization, and system acceptance testing by a purchasing jurisdiction:
>
> Qualification testing follows the vendor's developmental testing;
>
> Qualification testing provides an assurance to state election officials and local jurisdictions of the conformance of a voting system to the Standards as input to state certification of a voting system and acceptance testing by a purchasing jurisdiction; and

Qualification testing may precede state certification testing, or may be conducted in parallel as established by the certification program of individual states."

VSS Section 9.4.1.4 notes that:

> "The interface between the voting system and its users, both voters and election officials, is a key element of effective system operation and confidence in the system. At this time, *general standards for the usability of voting systems by the average voter and election officials have not been defined, but are to be addressed in the next update of the Standards*. However, standards for usability by individual voters with disabilities have been defined in Section 2.7 [sic: should be 2.2.7] based on Section 508 of the Rehabilitation Act of 1973, as amended in 1998. Voting systems are tested to ensure that an accessible voting station is included in the system configuration and that its design and operation conforms with these standards." [emphasis added]

Appendix C is the VSS's preliminary statement on requirements for usability. The requirements are a mixture of functional and design specifications, with the latter being somewhat predominant. The level and specificity of the requirements vary greatly. For example:

> "...the cursor should be automatically positioned in the first data entry field and when the voter hits the 'enter/return' key, the cursor should automatically move to the next data entry field;"

> "...fields where voters have to enter identifying information, if any, should be clearly labeled and the place where the information is to go should be clearly visible;"

> "The display should provide orientation and landmark features to support the voter in determining where they [sic] are in the ballot;"

Section C.1 emphasizes formative rather than quantitative/summative testing. For example, this section states:

> "Results from the tests and evaluations can be used to correct any design deficiencies before the system are [sic] actually used for voting."

4.1.2.3 Pertinent Features of VSS: Volume II, Testing Standards

Volume II provides a set of guidelines for conducting conformance tests. As with Volume I, some of the guidance is very specific and some is very general. Overall, Volume II seems to be a requirements document for the ITAs rather than the description of a specific test suite. However, the test methodology is left up to the ITA. VSS Section 1.5 gives ITAs the power to develop tests as needed, even if the areas are not directly covered by the VSS:

> "Taking advantage of the experience gained in examining other voting systems, ITAs will design tests specifically for the system design, configuration, and documentation provided by the vendor. Additionally, new threats may be identified that are not directly addressed by the Standards or the system. As new threats to a voting system are discovered,... ITAs shall expand the tests used for system security to address the threats that are applicable to a particular design of voting system."

This latitude in designing and conducting tests across voting products may be appropriate to allow the ITAs to develop specific tests based on the nature of the technology used, but would not ensure uniform testing of the independent quality of usability and accessibility across all voting products.

VSS Section A.4.3.5 (System-level test case design) tells the ITAs to simulate typical voter errors and gauge the robustness of the system, but this is not the equivalent of actual user interaction:

> "Usability tests: These tests are designed to exercise characteristics of software such as response to input control or text syntax errors, error message content ..."

VSS Section B.5 authorizes ITAs to use their own judgment to decide, in some cases, whether a system is accepted or rejected:

> "Of note, any uncorrected deficiency that does not involve the loss or corruption of voting data shall not necessarily be cause for rejection. Deficiencies of this type may include failure to fully achieve the levels of performance specified in Volume I, Sections 3 and 4 of the Standards, or failure to fully implement formal programs for quality assurance and configuration management described in Volume I, Sections 7 and 8. The nature of the deficiency is described in detail sufficient to support the recommendation either to accept or to reject the system, and the recommendation is based on consideration of the probable effect the deficiency will have on safe and efficient system operation during all phases of election use."

4.1.2.4 Current Conformance Testing

Currently, there is one laboratory accredited for hardware testing and two for software. The testing process is intensive and time-consuming. A "smooth" qualification process for both hardware and software testing generally takes several months. (A detailed description of the testing process may be found at: http://www.nased.org/ITA_process.htm.) Design guidelines supporting accessibility in the VSS are tested as part of this process (mainly by inspection). However, this current process does *not* include any significant usability testing or other conformance testing pertaining directly to usability. The VSS Appendix C is advisory and not mandatory.

4.1.2.5 Recent FEC Efforts in Support of Usability

The FEC organized an Advisory Board on Usability and Human Interface Standards in the Fall of 2002 to investigate how concerns about usability and interface standards could be incorporated into the VSS. More recently a series of usability guides for voting systems (FEC Developing, 2003; FEC Procuring, 2003; FEC Usability Testing, 2003) have been issued to assist state election officials and voting system vendors in the design, development, and procurement of usable voting products. While these brochures have high educational value as short tutorials, they do not provide specific details, procedures, or criteria necessary to determine if a product is usable.

4.1.3 IEEE Effort

The Institute of Electrical and Electronics Engineers (IEEE) has undertaken Project P1583 on Voting Equipment Standards under the Standards Coordinating Committee 38 (SCC 38) on Voting Standards to formulate standards for DRE voting equipment. IEEE charged this project as follows:

> "Project P1583 is charged with development of a standard of requirements and evaluation methods for election voting equipment. The standard will provide technical specifications for electronic, mechanical, and human factors that can be used by manufacturers of voting machines or by those purchasing such machines."

Details can be found at http://grouper.ieee.org/groups/scc38/1583/index.htm .

It appears that paper-based systems, such as optical scan systems, are not covered by their effort. The standard is currently (as of October 2003) in draft form and is undergoing review and editing. Within the standards, IEEE Section 5.3 addresses usability and accessibility issues The specifications are a mix of high and low-level, and performance and design requirements. Low-level design specifications predominate in the standard. Only voting equipment is covered in the body of the standard, but there is an informative annex offering guidance for ballot design.

IEEE Section 6.3, which covers testing, classifies testing methodologies into two categories: Standards Compliance and Usability Testing. In the Standards Compliance section, four different methods are used to determine whether or not a voting system conforms to each applicable usability/accessibility standard: inspection, expert –based evaluation, and tests and usability tests.

Inspection (I) the design is inspected to determine whether it possessed a feature or function specified in the standards...

Expert-based analytical evaluation (E) a human factors or usability subject-matter expert performs a comprehensive review ... to determine whether the applicable standards are being met.

Test (T) tests are specified to determine whether the applicable standards are met, e.g., a measure of letter height or sound intensity.

Usability testing (UT) evaluates a voting system by having a representative sample of voters perform voting tasks under realistic but simulated conditions. User performance and user opinion regarding their interactions with the system are measured and compared against usability and accessibility goals and requirements.

The various requirements in IEEE Section 5.3 are mapped onto one of these methodologies as the preferred way to verify that the system under test conforms. Usability testing is mainly reserved for general usability requirements that cannot be tested by the other compliance testing methods.

4.2 Generic Usability and Accessibility Standards

There are a number of standards for usability and accessibility. These standards are typically written to apply across large domains such as military systems, computer applications, or web site designs. As generic usability standards, they do not address functional issues, since they cannot account for the intended users, activities, and goals of a product being developed under these standards. In addition, as generic standards they do not include specific performance requirements, since such requirements also depend on the application domain.

These generic standards contain examples of the various kinds of requirements as described above in Section 2.3 (performance vs. design, specific vs. general, etc.). One further distinction is worth making: some of these standards apply to products in the conventional sense. For others, it is the development process itself that is specified. We refer to the latter as “process-oriented” standards.

Several of the standards are ISO standards for usability. The U.S. does not have anything equivalent as a national standard except for ANSI/HFES 100-1988 (HFES-100, 1988), which covers only ergonomic (physical) requirements of workstations in an office setting; it does not cover software interface design issues or processes. Several Military standards exist that address human factors engineering concerns for equipment (including software) and facilities, as well as for specific military applications (such as helicopter cockpit design), for specific items (such as labeling), as well as planning and process requirements. Many of these documents are no longer supported. One notable exception for process standards development is the Department of Defense, which has standards such as MIL-STD-1472 (MIL-STD-1472) and MIL-H-46855 (MIL-H-46855), covering human factors engineering. But, process documents like MIL-H-46855 and others are no longer being maintained and have been rescinded. Companies in the U.S. tend to rely on industry standards and best practices for their guidance with such documents as the Windows Style Guide (which changes with each release) and other commercially available books on interface design. Another exception is ANSI/INCITS 354-2001 (ANSI/INCITS 354, 2001) a standard

developed by NIST for documenting summative usability test results. It is currently in the process of internationalization. Some of the generic standards that are applicable to voting are described below.[15]

4.2.1 Section 508 of the Rehabilitation Act of 1973, as Amended in 1998

The U. S. Access Board is an independent Federal agency devoted to accessibility for people with disabilities. Part of its mission is to develop and maintain accessibility requirements for the environment, transit vehicles, telecommunications equipment, electronic and information technology, and to provide technical assistance and training on these guidelines and standards. It also enforces compliance for federal buildings under the Architectural Barriers Act. The Access Board developed the ADA Accessibility Guidelines for Buildings and Facilities (ADAAG, 2002) that apply to any building where a poll is located. The physical and environment accessibility requirements in the current VSS are based on these standards for clearances and reach, etc. Of particular interest for DREs are the Section 508 Electronic and Information Technology Accessibility Standards (Section 508, 2000). The Standard provides information on accessibility for operating systems and computer applications (including web sites) and it covers both self-contained, closed products and "open architecture" products. The VSS and the IEEE draft standards based their DRE accessibility standards on the Access Board Section 508 Standards.

As stated in the official Section 508 website (see: http://www.section508.gov):

> "In 1998, Congress amended the Rehabilitation Act to require Federal agencies to make their electronic and information technology accessible to people with disabilities. Inaccessible technology interferes with an individual's ability to obtain and use information quickly and easily. Section 508 was enacted to eliminate barriers in [Federal] information technology, to make available new opportunities for people with disabilities, and to encourage development of technologies that will help achieve these goals….Under Section 508 (29 U.S.C. ' 794d), agencies must give disabled employees and members of the public access to information that is comparable to the access available to others."

Subpart B contains the technical specifications that apply to a wide variety of IT products including software, web-based applications, multimedia and PCs. From the 508 summary page:

> "This section provides technical specifications and performance-based requirements, which focus on the functional capabilities of covered technologies. This dual approach recognizes the dynamic and continually

[15] It should be noted that many of the standards discussed in this section are long and complex and only a highly simplified overview is presented here. For greater detail, readers are urged to consult the standards themselves.

evolving nature of the technology involved as well as the need for clear and specific standards to facilitate compliance. Certain provisions are designed to ensure compatibility with adaptive equipment people with disabilities commonly use for information and communication access, such as screen readers, braille displays, and TTYs."

Thus, we see that the Standards encompass both design and performance specifications, and both quality-oriented and interoperability requirements. As with other standards we have examined, some of the requirements are very low-level and specific, others are very general.

4.2.2 ISO 9241: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)

The 17 parts of this standard cover many aspects of working with VDTs. It is one of the largest and most detailed usability standards in effect today. The various parts of 9241 exemplify the extremes of standardization. Some parts are highly technical and require very specific quantifiable properties. Other parts offer general guidance on how to approach certain tasks.

As an example of a technical specification, consider Part 7: Requirements for Display with Reflections. Section 4 carefully defines technical concepts and metrics and how they are related mathematically. This allows precise characterization of hardware performance (e.g. luminosity at various angles). Section 5 states that the purpose is to assure that VDTs be "legible and comfortable in use". Section 6 lays out the precise requirements to be met. Section 7 then describes the approved test method (lighting, optical instrumentation etc.) for ascertaining conformance. It is notable that the standard itself defines the test method as well as the requirements.

The next notable point is that the standard anticipates (as a future technique) a completely different test method based on the *performance of human subjects*, namely their ability to read text from the screen under various lighting conditions. This test method more directly addresses the purpose of the standard as set out in Section 5, but is less closely tied to the technical requirements of Section 6. Thus, there are really two approaches described within the standard:

- Specifying the physical characteristics of the light coming off the screen, as measured by optical instruments (in our terminology, a design, specific, low-level standard), and

- Specifying the resulting legibility of text and graphics, as measured by human performance testing (a performance, specific, high-level standard).

Later in the same document, in contrast with the highly technical nature of Part 7, Part 10: Dialogue Principles is a very general, non-binding standard. As its name implies, it simply describes some design principles that should be taken into

consideration when developing a system that interacts with human users via a VDT. It is a good overview and tutorial, but is not a standard in the strict sense. Likewise, Part 11: Guidance on Usability defines and discusses basic concepts of usability (effectiveness, users, tasks, etc). The discussion is very thorough - comparable to a long introductory chapter in a textbook.

4.2.3 ISO 13407: Human-centered Design Processes for Interactive Systems

This standard "provides guidance" to project managers on how to incorporate human-centered design into their development processes. The tone and content is more like a survey article, rather than a true standard (note the use throughout of the non-binding "should"). Like the Common Industry Format for Usability Test Reports (see below), this is a process-oriented standard.

In the conformance section, the standard indicates that the project manager must generate documentation showing that the procedures of 13407 were followed. The "level of detail" of the documentation is to be negotiated by the "involved parties". Annex C provides templates for such documentation, which is to be evaluated by an "assessor". Thus, the test procedures fall into the category of subjective inspection.

ISO 13407 contains a good deal of useful information -- it provides a checklist of possible techniques to be used by the conscientious project manager who wishes to improve the usability of a product.

4.2.4 ISO 16982: Ergonomics of Human-System Interaction -- Usability Methods Supporting Human-Centered Design

This document supplements ISO 13407. As a technical report (TR) it is "entirely informative" and is therefore not a standard. Its purpose is to describe several usability methods and recommend when they are most applicable. Thus, like ISO 13407, this document is basically advice for the project manager.

Twelve methods are discussed, eight of them involving the use of human subjects. Overall the TR presents a useful survey of available methods and techniques. As a consensus document, it is not "cutting edge" but rather collects, organizes and presents the common wisdom on applicability of usability methods.

4.2.5 ISO 10075: Ergonomic principles related to mental workload

This standard has two parts: General terms and definitions, and Design principles. Part 1 contains definitions and general terms. Part 2 contains a discussion of general design considerations and approaches. Included topics are: ambiguity of goals (more ambiguity implies more stress), complexity of tasks, and redundancy of information. Possible solutions include frequent rest periods, better illumination, and job rotation. It is not really a standard in the strict sense,

but rather a general discussion of some high-level problems and goals related to usability.

4.2.6 ANSI/INCITS 354-2001: Common Industry Format (CIF) for Usability Test Reports

The ANSI/INCITS 354-2001 defines a format to be used by someone who has conducted a usability test on a product (vendor or third-party) so that the report of the results of that test is reported in a standard and interchangeable way, thereby encouraging buyers to take these results into account as part of the total cost of ownership when procuring software. This is not a specific design standard, but one oriented towards the *process* of evaluating usability.

5 Current Human Factors Engineering, Usability, and Accessibility Research

This section summarizes research that can be applied to voting systems for both design and testing. Research can be divided broadly into basic and applied. First, as background, we discuss basic versus applied research. We then describe the types of results that can be applied to voting.

5.1 Background: Basic Research

Basic research is general in nature and therefore can be applied across a broad range of design domains (including voting). As a result, there is a great deal of data available from human factors, ergonomics, cognitive psychology, human computer interaction, usability engineering, and other related fields that are applicable to voting systems. This includes data on basic human perception, memory, cognition, higher level thought processes, decision-making, biases, psychomotor capabilities, etc. But basic research investigates one or just a few variables in isolation. As a result, basic research results can be difficult to apply within a given context such as voting system design. For example, basic research on human memory has been conducted to determine the number of unique elements that can be stored in human short-term memory. The study looked at colors, sounds, angles, or other simple data elements to be stored in human short-term memory to determine the actual limit.[16] Whether the results of this research apply to a specific voting system product design depends on the design's use of short-term memory. Finally, it should always be kept in mind that basic research results are applicable to the specific participants used in the study and might not be applicable across a broad range of users, such as the users of voting products.

In addition to issues of applicability based on the research design parameters used, there is the issue of the interaction effect of variables. By isolating a single variable of interest, research can make conclusions about this variable. However, in a real world situation, this variable may interact with other variables. For example, the current VSS standards include information on minimal size for text to ensure the text is readable. This value is based on research with font size as an isolated variable (i.e., with lighting, contrast ratio, color, font style, and other factors held constant). In a real world situation, differences in contrast ratio, font style, lighting condition, display density, or even user fatigue could change the minimal font size requirements. Under some conditions, the minimal

[16] This is an actual reference to the data used in a study done in the late 1950's on human short term memory by Miller (Miller, 1956) that led to the now famous7 +/- 2 rule. The results were widely applied and led to the creation of 7 digit telephone numbers.

requirements for font size might need to be higher than stated and under other combinations, lower font sizes might suffice.

5.2 Usability Research Related to the Design and Testing Process

One of the most promising research areas in human factors engineering and usability fields that can be applied to the development of highly usable voting systems is the research on the product design and testing process. Much of the most recently published literature has focused on the "user-centered design" process as a design approach that directly enhances product usability (Bittner, 2000, Constantine, 2003, Desurvire, Kondziela, & Atwood.1992, ISO 13407, 1999, Jacobsen, & Jørgenson, 2000, Meister, 2000, Mercuri, 2002, Redish, Bias, Bailey, Molich, Dumas, Spool, 2002). This process is described under a number of names and has been the venue of many consulting practices that specialize in business process reengineering as well as usability labs in large software development companies.

The user-centered design (UCD) process, and its derivative forms, is an approach that includes interaction with users throughout the product's design and development cycle to gather data and test design assumptions. The basic concept is to ensure that usability is incorporated into a product's design from the beginning of the design process and evaluated throughout the development process. Methods of incorporating usability include the use of user profiles (or personas), the development of use case models, usability walkthroughs, heuristic reviews, expert review, and user-based testing. Story boards, mock-ups, and prototypes can each be evaluated to test design assumptions and interaction effects. These activities serve to provide formative or diagnostic data on a product from conception through deployment.

Research is also being conducted on the proper use of test methods as part of the UCD process. This research examines the selection, application, and validity of various test methods typically used in a user-centered design process. But the UCD process itself is neither necessary nor sufficient to ensure usability. It is not strictly necessary in the sense that some teams have a good sense of design and may produce a fully usable product without formally adopting a user-centered design process. Conversely, having a UCD process in place does not guarantee a usable product, since the process must be applied by designers with understanding who can gather and apply the appropriate data. However, research does suggest that a user-centered design approach increases the *probability* of developing fully usable products.

5.3 Background: Applied Research

In contrast to basic research, applied research is more specific to a domain. There are applied studies from other domains that might be applicable to some extent, such as studies of ATM terminals, general computer system design, and

touch screen evaluations, but these studies suffer from the same limitation as basic research in their direct applicability and ability to be generalized to the total voting population. There is considerable research specific to ballot design issues and their effect on voting results in terms of bias (Alvarez, 2002; Darcy & McAllister, 1990; Design for Democracy (undated); Niemi & Herrnson, 2003; Roth, undated; Traugott, 2002). Unfortunately, our research uncovered little applied research in the area of design of equipment for voting products except for some small studies that lack statistical validity and reliability (Bederson & Herrnson, 2002; Bederson at al, 2002; CalTech-MIT, 2001; Englehardt & McCabe, 2001; Conrad, Unknown; Roth, undated; Roth, 1998; Tadayoshi et al, 2003). Similarly, our investigation was able to identify some limited research on accessibility issues associated with voting, but these studies were informal in nature and also lacked statistical validity and reliability (Burton & Uslan, 2002; Fields , 2003; Jones, 2002).

5.4 Usability Research Specific to Existing Voting Products

The studies that we were able to find on the specific issue of the usability of voting products would best be classified as formative. They provide findings and make specific recommendations or observations about specific products under evaluation. Some other studies have been based on data gathered from after-event reports or user opinion data, both of which may lead to false conclusions about both the nature and frequency of usability and accessibility problems, as we have previously discussed. These include reports such as the Caltech/MIT Report (Caltech 2001), the National Center for Voting Technology report (ECRI 1998), the New York Times story on voting results (McIntire, 2003), the University of Maryland report on the Diebold system (Bederson & Herrnson, 2002); the AccessWorld report on accessibility (Burton & Uslan 2002); product reviews by state officials as part of their product selection or evaluation process; and product reviews by end users and end user advocates such as the National Federation of the Blind and the American Foundation for the Blind.

Our review of these reports also suggests that many of the "results" of these studies were speculations about usability problems that *could* occur or *might have* occurred with the use of these products. For example, the University of Maryland expert evaluation of an early version of the Diebold DRE system in the State of Maryland identified multiple usability issues:

- No help button that can be used while voting
- No warning for overvoting
- The audio-only system was "hard to navigate", "difficult to have questions repeated", "poor audio quality", "no feedback after button presses"

The Caltech/MIT study speculated on usability issues by reporting on "spoiled ballot" data and making an estimate of the number of usability errors that were

likely represented by this number. It stated that 1.5 million presidential votes are lost each election and 3.5 million votes for governor and senator are lost each cycle. However, a null ballot may be the result of a machine failure to record the voter's preferences (a hardware or software problem), a voter error resulting from a usability problem, or an accurate record that the voter did not wish to vote for that office. However, all of these would be classified as "spoiled ballot" by the study's definition. Spoiled ballots strongly suggest usability problems that result in total failure. However, since the actual users were not interviewed or observed, there is no way to tell if usability issues were responsible for all of the spoiled ballots or if there is another cause (such as purposive voter action).

It is not clear that there is even a consistent definition of spoiled ballots across the studies or that any specific definition used is valid. For example, the New York Times article reported that official estimates stated that 60,000 votes were not cast in a 2000 election based on the lack of an interlock device on the voting product. It is not clear that the Caltech/MIT study, or others, would consider this a spoiled ballot.

In summary, to the best of our knowledge, there has been only one research effort [Roth, 98] performing a *controlled* experiment in which error rates for voting were *directly* measured by comparing the intended vote with the recorded vote. However, even this study did not include a fully representative sample of the voting population. Though the data from the Roth study is valuable, it was performed using 32 subjects, none of whom were users with disabilities, and did not attempt to provide specific tasks in an attempt to determine a range of potential usability problems that may be present in the product tested. Only a single ballot was used instead of a range of ballots. As a result, the data cannot be generalized across voters or ballot types for even the specific machine tested.

The data from informal reviews of voting products by officials and other parties interested in usability have raised similar issues related to the data they produce (Tutt.et al., 94, Etgen & Cantor, 2000, Gray & Salzman, 1998, Gray, 2003, Hertzman, et al., 2002, John & Marks, 1997). These studies have not been uniformly conducted in a realistic environment, with realistic ballots, or even representative users. Some of these studies are based on an evaluation process known to mask some usability problems and generate others as artifacts of the process. Some of these studies were "expert reviews" that report results based on the opinions of experts in the design of products. The results of these studies have been very important in raising awareness of usability issues and generating thought-provoking examples. However, it is likely that many of the identified usability problems would not change the election outcome (usability problems prior to success) and others might reflect problems that might not exist in actual use. The extent of these problems (how often they would show up in an actual election), and their actual effects are not known. Additional research is certainly required in this area; the recommendations presented in Section 6 address some of the limitations of existing research.

6

In this section we summarize our findings into ten recommendations based on the analysis discussed in this report. The recommendations focus on the need for an updated VSS that contains clear and unambiguous standards for usability and accessibility that are accompanied by conformance tests. These standards should not only reflect current research in human factors engineering, usability and accessibility but also make use of the best practices available for user interface design and standards specification, testing, and certification. We also must emphasize that the development of good standards is iterative and we would expect that it will take several years of development and supporting research to achieve these goals in their entirety.

We expect that these recommendations will be taken into consideration by the Technical Guidelines Development Committee (TGDC) when it becomes operational under the Election Assistance Commission (EAC) as described in the HAVA. Any implementation of recommendations will be at the behest of the EAC through the TGDC.

Please note that although a rationale is included with each of the recommendations in this section, we strongly suggest that the reader refer to earlier sections and the glossary in Appendix A for the more comprehensive assessment and analysis. In particular, section 2 defines and discusses many of the concepts used in the recommendations, e.g. performance vs. design requirements.

6.1 Overall Goal: Develop Measurable, Performance-Based Standards

6.1.1 Recommendation

Develop voting system standards for usability that are performance-based, high-level, and specific.

6.1.2 Rationale

Our assessment of the current application of standards to voting usability and accessibility is driven by the following design philosophy. We assume that the major goals to be promoted by usability standards for voting are the "classic" ones for usability, as described in ISO 9241-11, and discussed in Section 2.2.3:

- Effectiveness (e.g., voter votes for intended candidate)
- Efficiency (voter completes voting task within reasonable amount of time and effort)

- Satisfaction (voter's experience is not stressful)

For voting systems, this can be summed up as follows: A voting system is usable if voters can cast valid votes as they intended, easily and efficiently, and feel confident about the experience. Because we have defined "usability" to include usability by people with disabilities (see Section 2.2.2), the same measures apply for accessibility, once the barriers to accessibility are removed via a separate set of design standards to make a system available to those individuals.

These goals are, of course, those of *quality-oriented* standards, not interoperability or metric-based standards. Other things being equal, such standards are best formulated with these properties:

- Performance (not design): because the goal concerns the functions and sub-functions that the system is capable of doing, not the mechanism by which the functions are supported;

- High-level: because it is necessary to specify only the basic operations intrinsic to the entire application, not to give details about the underlying sub-functions; and

- Specific (not general): because conformance tests should be objective and justified, not subjective or arbitrary.

Such standards, and the conformance tests based on them, directly address the bottom-line performance of existing products. They do not attempt to guide product development, nor diagnose problems. Further, this approach is supported by the ITA structure currently in place. The process the ITAs use to certify a voting system is based on testing against a standard. As such it is critical to have standards that lend themselves to objective, repeatable and reproducible test procedures.

By these criteria, many of the usability specifications in the VSS (VSS Section 3.4.9 and Appendix C) and in the IEEE draft standard suffer from one of two problems. They are either too design-oriented and low-level, or too general (even when they are performance-oriented). So, while the existing and draft standards are a good base, they need some reformulation.

6.2 Specify Functional Requirements

6.2.1 Recommendation

Specify the complete set of user-related functional requirements for voting products in the voting system standards.

6.2.2 Rationale

Though the current VSS do contain information on functional requirements, it is unclear that the functional requirements specified are complete and, as written, they can be interpreted in more than one way. A lack of precise definition of functional capability can significantly affect usability and accessibility. For example, recent discussions have centered on the audio playback of candidate names. According to a representative of the National Federation of the Blind, they had made a request that the DRE equipment they reviewed include the ability for the user to skip the automatic presentation of candidate names. According to this representative, the vendor stated this was technically possible, but they had been informed by the voting officials that full presentation of all names was a requirement for voting equipment. The ramification of this interpretation in terms of time on task and user satisfaction is obvious, particularly in extreme cases such as the recent California recall election where there were more than 100 names to be read.

The functional requirements should be at the user interface (not with the internal software requirements), should be independent of the implementation (make no references to "how", just "what"), and should not include imprecise references to "how well"(including metrics). The requirements should include the identification of the system level capabilities (e.g., voting for only one person in a single-seat contest, voting for multiple people in a multi-seat election, etc.) as well as the sequence control functionality of the interface (e.g., provide a means of moving between pages in a multi-page ballot, provide a means of moving between contests and/or referendums, etc.). This is not a simple effort, and indeed may be tedious in some respects, but it is absolutely necessary to address the specific functional requirements if we are to have appropriate usability guidelines. The requirements should also address ballot design software capabilities. The data necessary to specify these functional requirements could be derived from a complete human factors task analysis and could be validated as part of usability test development as described later in this section.

6.3 Avoid Detailed Product Design Specifications for Usability

6.3.1 Recommendation

Avoid low-level design specifications and very general specifications for usability. Only those product design requirements that have been validated as necessary to ensure usability should be included as "shall" statements in standards.

6.3.2 Rationale

As discussed in detail earlier in this report a number of issues are associated with the inclusion of detailed product design specifications in a standards document. The design specifications currently in the VSS and draft IEEE standards range from general statements (e.g., "A clearly legible font should be utilized") to tables

of possible options and formulas (e.g., a list of approximate point sizes for text based on anticipated viewing distances), to specific requirements. Design is an active process and design data must be applied and then validated to ensure the end result is what was intended or desired by the designer (i.e., that usability or accessibility is maintained or enhanced). The background and expertise of the audience need to be considered in determining the nature and scope of the guidance provided. Finally, maintaining the most appropriate design guidance in a standard is an ongoing issue as new technology; changes in existing technology; and advances in our understanding of human psychology and decision-making, human-computer interaction, and ballot design continue to evolve.

Since the inclusion of specific design requirements appears to be part of the current approach for both the VSS and proposed IEEE standards, a more detailed discussion of the problem of providing detailed product design specifications is included below. This should not be viewed as an attack on the existing standards or the existing approach, but an assessment of the difficulties and limitations inherent in this approach. And, we believe there is a viable alternative in the development of conformance tests for usability and accessibility, as discussed below in Section 6.10.

6.3.2.1 Low-Level Design-Oriented Specification

As an example of specifications that are too design-oriented and low level, Section C.6 (i.) of the VSS reads:

> "In computer-based systems, the cursor should be automatically positioned in the first data entry field, and when the voter hits the "enter/return" key, the cursor should automatically move to the next data entry field."

This might be a good design guideline, but a standard should not so closely mandate the design of a system. If cursor control is indeed a problem, it will be reflected in one or more of the basic usability measures (effectiveness, efficiency, or satisfaction). A long list of design guidelines, however valid they are individually, does not constitute a good standard. Note that we are *not* questioning the value of design guidelines, as such. These may be very helpful during the design and development of a voting product, but they are not *essential metrics* by which potential purchasers should judge the system. In addition, this specification presumes that a computer-based DRE has a cursor and an "enter/return" key. Depending on the design, a touch screen, for example, might not have these artifacts. Also, a list of design guidelines, whether provided as a standard or not, raises the questions of the validity and completeness of the list of requirements. Has it been shown that automatic cursor positioning actually does improve speed or accuracy in the voting process? Are there other equally valuable guidelines that have been omitted (e.g. adequate spacing between buttons)? Is there an interaction effect between guidelines that could affect the

specifics (e.g., contrast ratio and font height)? Are these interaction effects taken into account?

Problems with such a standard are also reflected in the testing process. A long list of low-level guidelines invites a long "checklist" or even "decision tree" style test to see if the requirements are met. This is a tedious process and does not ensure usability.

Further, in addition to the problems already noted, many of the current requirements are stated as ones that "should" be applied. This is non-binding so the vendor is not required to conform and it cannot be included in conformance testing. Because "guidance" is not enforceable, it is unclear that any *product* design guidance provided would ensure usability of the product. It would seem to be more appropriate to provide a discussion about the need to locate and apply the most current design guidance available, and the standard could also identify some of the more likely sources of such information.

Finally, as mentioned earlier, the IEEE draft standard addresses only DRE equipment and not paper-based systems, such as optical scan, and so could not serve as the basis for a general standard on voting and usability. Section 301c (2) of HAVA explicitly states that paper-based voting systems are not excluded. The standards provided to vendors should be applicable to any product they develop.

6.3.2.2 Imprecise Specifications

As an example of a specification that is too general, consider the following from Section C.4 (b.) of the VSS):

> "[ballot] Instructions [to the voter] should be concise. Instructions should be designed to communicate information clearly and unambiguously so that they can be easily understood and interpreted without error."

As a general design guideline to developers, this is unobjectionable, but as a specification it fails to provide a clear criterion against which conformance can be measured. Generally worded specifications have to be tested either by invoking an expert's judgment, which can be subjective, or by "creatively" interpreting the specification so as to generate a more precise test.

6.4 Address the Lack of Specific Research on Usability and Accessibility for Voting Systems on Which to Base Requirements

6.4.1 Recommendation

Build a foundation of applied research for voting systems and products to support the development of usability and accessibility standards.

6.4.2 Rationale

As discussed earlier, much of the human factors research is basic or is applied to an isolated variable. The interaction and additive effects across various variables are difficult to assess, and all of the data needs to be assessed in the specific domain of voting before it can be legitimately included.

Until very recently there has been little applied research from the human factors and usability fields specifically on voting systems. Accessibility has been addressed by generic design standards that intended to remove barriers to access, but usability by persons with disabilities has not been addressed by research. In fact, we know very little about users' experiences with voting systems including those people with disabilities. This suggests a need to focus efforts on building a foundation of applied research for voting systems and voting products to support the development of usability standards. Until this is done, there is little basis upon which to include many detailed specifications.

6.5 Develop Design Specifications for Accessibility

6.5.1 Recommendation

To address the removal of barriers to accessibility, the requirements developed by the Access Board, the current VSS, and the draft IEEE standards should be reviewed, tested, and tailored to voting systems and then considered for adoption as updated VSS standards. The feasibility of addressing both self-contained, closed products and open architecture products should also be considered.

6.5.2 Rationale

To properly address unaided use of voting machines by persons with disabilities, the Federal standards must address the removal of physical and cognitive barriers to accessibility. Design specifications, as described earlier in this report, must be unambiguous requirements that are to be met by the vendors. However, valid requirements should also state the intended effect in the product design, that is, state which usability or accessibility issue is addressed or what barrier to accessibility is removed. Since the interaction effect between specifications must be investigated, it is particularly difficult to state unequivocal design specifications. Despite these difficulties, some design specifications can and should be provided, particularly in the area of external physical requirements. These would address such issues as button spacing, force requirements, display and control surface angles, and reach distances. Many of these requirements are driven by issues of accessibility and are currently covered in the VSS and the IEEE draft standard.

In contrast to our recommendation for performance-based standards for usability, we believe that for accessibility, design standards are currently the only practical approach. This is because the population addressed by accessibility standards is so much more heterogeneous than that addressed by usability standards. As a

consequence, it is not practical to formulate performance criteria and test methods that could be applied broadly and uniformly to the disabled population.

The Access Board has provided information to the FEC for incorporation in the standards, which are the basis for the requirements currently in the VSS; however, the guidelines provided by the Access Board are for self-contained, closed products. These are products that are expected to contain all the accessibility features necessary for use by persons with disabilities. This is contrasted with "open architecture" products for which the end user is intended to provide some form of adaptive technology (e.g., a screen reader or external braille display). In addition, some of the requirements provided by the Access Board are general in nature and have not been tailored to the specific domain of voting system products. Also, they do not address all of the associated aspects that need to be specified (e.g., determining the quality of audio and how to test it) because some of this is considered usability rather than accessible design. The IEEE has made some progress in this area and any new VSS should take advantage of their work.

6.6 Develop Ballot Design Guidance

6.6.1 Recommendation

Develop ballot design guidelines based on the most recent research and experience of the visual design communities, specifically for use by election officials and in ballot design software.

6.6.2 Rationale

Ballot design is a complicated issue. On one hand, a great deal of flexibility is required in the design of ballots to allow election officials to adapt to specific elections. However, a large proportion of the usability issues reported to date relates to the design of ballots, and a significant amount of work has been done on the proper design of ballots – both for usability and to avoid bias. In addition to layout issues (including the number and arrangement of candidates on a ballot, avoiding misreading issues associated with tabular data, etc.), it appears that usability can be significantly affected by the wording of instructions.

We recommend that ballot design guidelines be developed based on the most recent research and experience of the visual design communities (e.g., the American Institute of Graphic Arts (AIGA)). These new guidelines would be for different audiences than the functional usability system requirements since ballot design involves vendor ballot design software and the users would be district, state, or regional government election officials, ballot designers and, possibly, printing vendors.

It is recommended that this be a separate section of the new VSS: a section that would include recommendations for ballot instructions, visual design and layout, and recommendations for randomizing. We believe that research is needed to

find an “optimal” set of guidelines, but significant improvement in usability could be made through standardization, particularly of instructions. This research would need to include developing instructions in alternate languages, since direct translation is not always possible and improper translations could induce new usability issues. In addition, we recommend that guidance be provided on testing methodologies to be used to ensure that the ballots do not inadvertently induce usability problems. Finally, it is recommended that a set of requirements to support these guidelines be developed and included as the specifications for vendor-developed ballot design software.

6.7 Develop Facility and Equipment Layout Guidance

6.7.1 Recommendation

Develop a set of guidelines for facility and equipment layout; develop a set of design and usability testing guidelines for vendor- and state-supplied documentation and training materials.

6.7.2 Rationale

Proper design of equipment must take into account the environment in which the equipment is to be used. This is true in any product design, but is a particular issue in voting product design due to the varying locations in which the products are used. There is an obligation on the part of election officials to operate the equipment in a suitable environment. These environmental factors include lighting, noise, temperature, equipment spacing, and a range of other elements. Though the voting officials do not have full control over these elements, too great a variation could induce usability and accessibility problems into an otherwise usable and accessible product. In addition, human performance can be affected by dealing with long lines. By anticipating the length of time that would be required with the given voting product, election officials can address the situation through predetermining that a sufficient number of voting stations are available as a means of dealing with long lines. In any case, we believe that a significant set of recommendations could be developed in this area that would enhance the overall experience and minimize usability and accessibility issues.

It is recommended that existing data be gathered and analyzed and a set of guidelines be developed for facility and equipment layout. This portion of the new standards would be for users other than vendors (i.e., election officials responsible for voting locations and poll workers). It would provide information on which the vendors could base their designs. Information relevant to facilities and equipment layout is very likely to be available in the research literature and can be generated almost entirely from a literature search.

We should not overlook the importance of the poll workers and election officials being able to set up the polls and run the election with the equipment properly. The usability of the documentation and training materials supplied by both the vendor and the state is critical. We recommend that these materials undergo

usability testing and that guidance be developed for how to do this testing at the state level.

6.8 Encourage Vendors to use a User-Centered Design Process

6.8.1 Recommendation

Encourage vendors to incorporate a user-centered design approach into their product design and development cycles including formative (diagnostic) usability testing as part of product development.

6.8.2 Rationale

As noted earlier, research and industry best-practices suggest that usability and accessibility are optimally addressed from the inception of a design and development process. From interviews with vendors, we believe many do not currently follow a user-centered design approach or conduct many of the activities specifically intended to address usability and accessibility in their design and development processes. Though they cannot be compelled to use a specific approach, nor do we believe that following such an approach will be sufficient, significant improvements in both usability and accessibility would likely result if they followed a user-centered design approach and if they conducted the activities specifically intended to address usability and accessibility. The FEC recognized this when it produced its brochure on the user-centered design process (FEC Developing, 2003).

We recommend that vendors be encouraged to incorporate a UCD approach into their product design and development cycle including formative (diagnostic) usability testing as part of product development. As the Federal standards are revised to incorporate more usability requirements, this will help vendors prepare for usability qualification testing. Further, we recommend that vendors be encouraged to perform their own summative usability testing on their products prior to releases and report them using a CIF standard (INCITS 354-2001) format.

6.9 Create Test Procedures for Accessibility

6.9.1 Recommendation

Develop a uniform set of procedures for testing the conformance of voting products against the applicable accessibility requirements.

6.9.2 Rationale

As stated above, it is recommended that the accessibility requirements be developed to remove the barriers to access. We believe there is sufficient data available (or can be generated) to develop a complete set of specific requirements for removing the barriers to accessibility.

We recommended that a uniform set of test procedures be developed for testing the conformance of voting products against the applicable accessibility requirements (self-contained, closed or open architecture products). Further, we believe that the test procedures could be added to the test battery currently conducted by the ITAs.

6.10 Create Test Procedures for Usability

6.10.1 Recommendation

Develop a valid, reliable, repeatable, and reproducible process for usability conformance testing of voting products against the standards described in the recommendation in 6.1.1 with agreed upon usability pass/fail requirements.

6.10.2 Rationale

In general, the **single most critical need** identified in this report is a set of usability standards for voting systems that are performance-based and that support objective measures and associated conformance test procedures that can be used for the certification and qualification of voting products and systems.

These measures and conformance test procedures primarily depend upon having the usability testing process recommended in this section.

As described in this report, we have separated accessibility into two categories: removing the barriers to access and usability by users with disability. Removing barriers to accessibility has already been addressed in the recommendations above and identified as a likely extension to the current ITA process; therefore, we will restrict our discussion in this section to usability. Note, however, that when we are referring to usability we include all users, with and without disabilities, at different levels of reading proficiency and from different cultural and economic backgrounds.

As we have discussed previously, a set of design requirements cannot properly address the issues of usability for voting system products. Also, no document can contain a sufficient set of design requirements to ensure voting product usability unless the document completely specifies a design already shown to be usable. And, the traditional ITA test approaches such as testing by demonstration or inspection will fail to uncover usability problems. The FEC brochures, the IEEE discussions on usability evaluation, and numerous reports and texts discuss testing that should be done as part of the design process. Though these formative or diagnostic tests are valuable tools in the design process, they do not guarantee that the final product is usable as measured by the metrics described earlier (efficiency, effectiveness, and satisfaction) since they are used during the design process, not on a final product. Even tests that are conducted on a final product design are generally not conducted in a way that would allow the results to be generalized to the intended populations (i.e., the participants of the study may or may not be appropriately extrapolated to a

majority of all actual users). This is particularly true for voting system products since the range of users required for such a test would make this type of testing cost prohibitive to most vendors. In addition, there are currently no defined standards for usability metrics that vendors could use as benchmarks for their testing. For these reasons, we believe that vendor testing of the product, while valuable, is a separate issue from certifying that the end product is usable. We believe that usability qualification testing is necessary, but it will require the establishment of both objective usability test procedures and pass/fail criteria.

To ensure usability of a voting product, it is imperative that the product be tested with actual users performing realistic tasks in a realistic environment, in sufficient numbers, and using a broad enough cross-section of users to be truly representative of the voting population. Further, to ensure good usability of the system, we must test not only the interaction of voter with the product but also the interaction of the voters, election administrators, and poll workers with the entire voting system.

We recommend the development of a valid, reliable, repeatable, and reproducible process for usability testing of voting products against agreed-upon usability pass/fail requirements. In particular, there must be a careful definition of the metrics, such as what counts as an error, how to measure error rate, time on task, etc., by which systems are to be measured. The pass/fail criteria should be restricted to usability problems leading to partial failure, and usability problems leading to total failure. Since we are dealing with outcomes, usability problems prior to success need not be specifically included, but would be represented in the time on task measure from testing. Note that while excessive time required does not lead to failure, it is still unacceptable.

Since human users are involved in the process, it is unlikely that the error rates will be zero for *any* criteria established, so a specific acceptable error rate and margin of error will likely be required. For example, it may be possible to enforce a requirement that no user be allowed to consciously cast a ballot with an overvote for one or more contests since this error represents the *action* of the voter. However, a voter still might inadvertently cast a vote for an unintended candidate in any product but this error cannot be detected without knowing the *intent* of the voter. Yet, both of these conditions must be tested. This test process must be defined at a high enough level of generality that the same procedure could be applied to any product (i.e., we do not want to define product-specific tests). Otherwise, the results for various products would not be comparable. Fortunately, the task requirements for voting are specific enough that this should not be difficult to do. It might be necessary, however, to have technology-specific variants of the test procedure and protocol (e.g. DRE vs. paper-based), although we believe the differences can and should be kept minimal.

Research would need to be conducted to determine: (1) the nature of errors possible during a voting process (this includes voter errors and poll worker errors), and (2) the level (rate) of these errors (both the current levels for existing products and recommendations for "acceptable" levels of each error type). Once this information is available, we recommend that a set of repeatable and reproducible processes be defined and that each voting product be tested using these test processes and usability test pass/fail criteria. This would include the definition of all test procedures, the data collection required, the data analysis approach, participant screening and selection procedures, and reporting requirements. We also believe that, though the ITAs would likely have the responsibility to conduct these tests, the nature and format of the testing would likely require additional personnel with qualifications to conduct this type of testing.

As part of the development of this report we have explored the feasibility of this recommendation and have provided some suggestions as to how to develop the test procedures and protocols. This information is included in Appendix B of this report. The details of the statistical data analysis are described in Appendix C.

7

7.1 Proposed Timeline

In this section we outline the initial steps needed to implement the recommendations we have suggested. In particular, developing a set of performance-based usability standards and associated test procedures is a complex endeavor. Also, even gathering together the existing standards, checking their validity, and ensuring that the ITAs have proper test procedures will require some effort. We also recognize that vendors are developing their products and state and local officials must make procurement decisions in the short term. Therefore, we also describe a preliminary roadmap for implementing these recommendations that includes suggestions for short-term activities that will help to address usability and accessibility issues while the longer-term research and development proceeds.

7.2 Short-Term

In the short term, we recommend a push to obtain initial user testing data as soon as possible. We anticipate this being a "pilot-test" with both disabled and non-disabled users to simply find the usability issues, and determine possible procedures for testing. We would simplify by using only 1-3 different ballots. This initial data will be used to develop robust testing protocols including appropriate statistical analyses. (We discuss the statistical analyses in Appendix C.) The initial data we gather from testing with real equipment would be forwarded to vendors so they can improve their products as they see fit.

We recognize, however, that some states are facing purchasing decision deadlines for products for the 2004 election and that they want to make wise choices that include usability and accessibility factors. We recommend the following for the state election directors:

- Ask vendors for a report of their summative usability testing preferably in the standard format of ANSI/INCITS 354 Common Industry Format for Usability Test Reports. Any data on formative usability testing would also be helpful.

- Ask a usability professional to

    - Conduct an evaluation

    - Interpret the vendor's tests reports, and

    - Evaluate the usability and accessibility of the voting products under consideration and test with typical past ballots using the usability

professional's choice of usability evaluation methods (which might not be user testing)

- After procurement, ask a usability professional to evaluate the voting system with actual ballots, voters, and poll workers before elections. If a usability or accessibility problem is identified at that point, stopgap remedies often exist, such as additional instructions for the voters or poll workers.

- After procurement, also ask a usability professional to evaluate the usability of any documentation for poll workers and election officials.

We recommend that vendors:

- Begin to implement a user-centered design approach in their system engineering process to prepare for ITA usability testing. The IEEE P1583 draft standard provides advice on this as well as design guidelines

- Document any usability evaluation and usability testing that products have undergone

- If possible, perform a summative evaluation of their products and report in the CIF.

Because of financial concerns and convenience, one or more states may plan to join together for these evaluations. The FEC has prepared brochures about usability and procurement of voting systems that are helpful in indicating the issues that should be considered before making a decision. Additional information on accessibility can often be gathered from various advocacy groups for the disabled, such as the National Federation of the Blind, United Cerebral Palsy, etc. who are often willing to do reviews of products, typically focused on one type of disability. Note that many of the results from this type of testing are subjective and often include usability problems prior to success and, therefore, should be used with caution.

7.3 Long-Term Plans – 1-4 Years

In the longer term, as part of a major effort, work should begin on the formulation of standards for usability, as discussed in recommendation 6.1 and, in parallel, on the development of standardized test procedures, which should be checked for validity and reliability, etc. The goal is to develop a set of validated procedures for the task based testing within 1-2 years. This set of procedures could be used by vendors, procuring offices, etc. Baseline performance levels could then be determined through applying these procedures on existing voting products in following 2 years. Obviously, the procedures themselves would not provide enough for an ITA process until error limits and confidence rates were confirmed.

7.4 Coordination with the TGDC

We expect that the recommendations in this report will be taken into consideration by the EAC and the TGDC. NIST will work with the TGDC to develop a plan for implementing the recommendations. In the next section, we suggest some basic work that we believe is needed to support implementation of the performance-based standards described in this report as we view this as critical to any VSS updates. We leave the details of a work plan and timeline for the other aspects of the recommendations to the TGDC. In general, work should also begin on the issues of functionality, ballot design, and facility layout, some of which will be to determine what research materials are already available.

7.5 Proposed Next Steps for Testing and Standards Development

To move forward to meet the goals of improving usability and accessibility of voting systems and products, the next steps have two major emphases: (1) initial baseline testing of voting products that are currently in use as a means to determine errors, procedures, and statistical criteria (based in part on the preliminary short term research project) 2) to develop the standards as described in this paper for usability and accessibility.

7.5.1 Proposed Testing

For the testing portion of this effort, we suggest these steps:

- Perform investigative studies to document the various types of voter errors that are possible on voting machines. There will be a baseline assumption from the short term work, so this effort would entail validating the assumption and modifying the baseline as needed. This should involve a "reasonable" number of participants (including both disabled and non-disabled people) but the participants need not be representative in any way.

- Develop the test procedures for objectively measuring the errors and performance. This will involve defining what constitutes a single error, defining how to collect timing data, developing the number of procedures to be used, selecting the procedure order, determining the specific wording for the procedures, developing the criteria for each element measured (time, objective data, subjective data), as well as creating supporting materials (training materials, sample ballots, moderator scripts for various things, etc.)

- Conduct studies to validate the procedures and check them for validity and reliability

- Perform a pilot study to determine current system baselines for existing voting machines

- Recommend error limits and confidence rates
- Determine sample rates and sampling methods
- Optimize the approach for costing (i.e., investigate ways to minimize the cost of conducting testing). This could involve the use of the Wald formula (see Appendix C) to reduce the number of participants needed, recommendations for simultaneous testing of machines, testing with populations most sensitive to detecting specific issues, etc.)
- Dry run the full test and hand off to the standards development process.

7.5.2 Proposed Standards Development

In parallel with the development of testing methods and collection of baseline data, a process for creating the next generation of VSS incorporating these performance-based standards should be developed. Again, the development of this process must be under the advisement of the TGDC.

References

ADA Accessibility Guidelines for Buildings and Facilities (ADAAG) (2002). http://www.access-board.gov/adaag/html/adaag.htm

Adelman, L. (1991). Experiments, Quasi-Experiments, and Case Studies: A review of Empirical Methods for Evaluating Decision Support Systems *IEEE Transaction on Systems, Man, and Cybernetics*, Vol. 21, No.2 p. 293-301

Alvarez, R.M. (2002). Ballot Design Options, Manuscript prepared for Human Factors Research on Voting Machines and Ballot Design: An Exploratory Study

Andre, T.S.; Belz, S.M.; McCreary, F.A.; & Hartson, H.R. (2000). Testing a Framework for Reliable Classification of Usability Problems *Proceedings of the IEA 2000/HFES 2000 Congress*, p. (6)573-576

ANSI/HFES 100 (1988). Human Factors Engineering of Visual Display Terminal Workstations

Bailey, R.W. (2000). The Usability of Punched Ballots: Improving Usability in America's Voting Systems, *Human Factors International*

Bederson , B.B. & Herrnson, P.S. (2002). Usability Review of the Diebold DRE System for Four Counties in the State of Maryland. Report from the University of Maryland's Center for American Politics and Citizenship, http://www.capc.umd.edu/rpts/MD_EVoteMach.pdf

Bederson , B. B. & Herrnson, P.S. (2002). An Evaluation of Maryland's New Voting Machines - CAPC and HCIL exit poll research on voter comfort and trust in new electronic voting machines, *A Report from the University of Maryland's Center for American Politics and Citizenship web site*

Bederson , B.B.; Herrnson, P.S.; & Niemi, R.G. (2002). Electronic Voting System Usability Issues, *A Report from the University of Maryland's Center for American Politics and Citizenship web site*

Bevan, N. (2000). ISO and Industry Standards for Usability Measurement, Tutorial Notes, SERCO

Bittner, A.C. Jr. (2000). Building Performance Measurement into Today's Testing and Evaluation (T&E). *Proceedings of the IEA 2000/HFES 2000 Congress*, (6) p. 557-560

Bremer, J. (Undated). Ballot Design in an Electronic Environment – Lessons from the Online Market Research Industry

Burton, D & Uslan M. (2002, November). Cast a Vote by Yourself: A Review of Accessible Voting Machines, Access World, 3(6), November, *http://www.afb.org/afbpress/pub.asp?DocID=%20aw030603*

CalTech-MIT (2001). July, Voting: What is, What Could Be, http://web.mit.edu/voting/

Castillo, J. C. & Hartson, H.R.; (2000). Critical Incident Data and their Importance in remote Usability Evaluation, Proceedings of the IEA 2000/HFES 2000 Congress, p. (6)590-(6)601

Cherlunick, P.D. (2001). Methods for Behavioral Research: A Systematic Approach, *Sage Publications, Inc.*

Conrad, F. G (Unknown). Usability and Voting Technology, *White paper for Voting Technology Workshop*

Constantine, L. (2003). Testing… 1… 2… 3… Testing… (unpublished)

Darcy, R., & McAllister, I. (1990). Ballot Position Effects. *Electoral Studies*, 9(1), pp. 5-17

Design for Democracy Case Studies (undated). From the *Design for Democracy* Web Site, http://www.electiondesign.org/case.html

deSouza, Flavio and Bevan, Nigel (1990). The Use of Guidelines in Menu Interface Design: Evaluation of a Draft Standard. *Proceedings of IFIP INTERACT '90: Human-Computer Interaction*, p. 435-440

Desurvire, H.W. Kondziela, J.M. & Atwood, M.E. (1992). What is Gained and Lost when Using Evaluation Methods Other than Empirical Testing Practical Evaluation Methods for Improving a Prototype *Proceedings of the HCI'92 Conference on People and Computers VII 1992 p.89-102*

Dutt, A. Johnson, H. & Johnson, P. (1994). Evaluating Evaluation Methods Methodology of Interactive Systems Development *Proceedings of the HCI'94 Conference on People and Computers IX 1994 p.109-121*

Englehardt, J. & McCabe, S. (2001, March 11). Over-votes Cost Gore the Election in FL, *Palm Beach Post*, http://65.40.245.240/voxpop/palmpost.htm

Etgen, M. & Cantor, J. A, (2000). Comparison of Two Usability Testing Methods: Formal Usability Testing and Automated Usability Logging, Proceeding of the 2000 UPA Conference

Federal Election Commission (2003). Developing a User-Centered Voting System, http://www.fec.gov/pdf/usability_guides/developing.pdf

Federal Election Commission (2003). Procuring a User-Centered Voting System, http://www.fec.gov/pdf/usability_guides/procuring.pdf

Federal Election Commission (2003). Usability Testing of Voting Machines, http://www.fec.gov/pdf/usability_guides/usability.pdf

GAO (2001). Elections: Status and Use of Federal Voting Equipment Standards, GAO-02-52, October, 2001, http://www.gao.gov/new.items/d0252.pdf

Gray, W.D. & Salzman, M.C. (1998). Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods, *Human-Computer Interaction*, *Vol. 13, p. 203-261*

Gray, W.D. & Salzman, M.C. (1998). Repairing Damaged Merchandise: A Rejoinder, *Human-Computer Interaction*, *Vol. 13, p. 325-335*

Gray, W.D. (2003). Returning Human Factors to an Engineering Discipline: Expanding the Science Base through a New Generation of Quantitative Methods – Preface to the Special Edition, *Human Factors*, Vol. 45, No. 1, p.1-3

Hemenway, D. (1980, October). Performance vs. Design Standards, NBS/GCR 80-297

Henninger, S; Haynes K.; & Reith M.W. (1995). A Framework for Developing Experience-Based Usability Guidelines, *Proceedings of DIS'95: Designing Interactive Systems: Processes, Practices, Methods, & Techniques,* p.43-53

Herrnson, P.S.; Niemi, R.G.; & Richman (undated). Characteristics of Optical Scan and DRE Voting Equipment: What Features Should be Tested? *http://www.capc.umd.edu/rpts/MD_EVote_HerrnsonNiemi.pdf*

Hertzum, M.; Jakobson, N.E., & Molich, R. (2002). Usability Inspection Methods by Groups of Specialists: Perceived Agreement in Spite of Disparate Observation, *CHI 2002*

ISO/TS 16071 (2003). Ergonomics of Human-System Interaction -- Guidance on Accessibility for Human-Computer Interfaces

ISO 9241-11 (1998). Ergonomic Requirements for Office Work with Visual Display Terminals (VDT) – Part 11: Guidelines on Usability

ISO 13407 (1999). Human Centred Design Processes for Interactive Systems

Ivory, M.Y. & Hearst, M.A. (2001). The State of the Art in Automated Usability Evaluation of User Interfaces *ACM Computing Surveys*, Vol. 33, No. 4, p. 470-516

Jacobsen, N. E. & Jørgenson, A. H.; (2000). The State of Art in the Science of Usability Evaluation Methods: A Kuhnian Method, Proceedings of the IEA 2000/HFES 2000 Congress, p. (6)577-(6)580

John, B.E. & Marks; S.J. (1997). Tracking the Effectiveness of Usability Evaluation Methods Usability Evaluation Methods *Behaviour and Information Technology*v.16 n.4/5 p.188-202

Jones, D. (2002). Handicapped Accessible Voting, *Voting and Elections Web Pages, University of Iowa*

Jones, D. (2002). Voting Systems Standards: Work that Remains to be Done, *Testimony before the Federal Election Commission, Washington D.C., April 17, 2002*

Jones, D.W. (2001). Problems with Voting systems and the Applicable Standards, *Testimony before the U.S. House of Representative*

Kanis, H. & Arisz, H.J. (2000) How Many Participants: A Simple Means for Concurrent Monitoring. *Proceedings of the IEA 2000/HFES 2000 Congress*, p. (6)637-(6)572

Leahy, M. & Hix, D. (1990). Effect of Touch Screen Target Location on User Accuracy, *Proceeding of the Human Factors Society 34th Annual Meeting*, p. 370-374

Lowgren, J. & Tommy Nordqvist, T (1992). Knowledge-Based Evaluation as Design Support for Graphical User Interfaces Tools and Techniques, *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems, p.181-188*

Lowgren, Jonas and Nordqvist (1992). Knowledge-Based Evaluation as Design Support for Graphical User Interfaces. *CHI '92*, p. 181-188.

McCormack, C.B. (2003). Ballot Design: Has It Impacted Voting Behavior in Los Angeles County, California? *Presentation at the 2003 CHI Conference*

McIntire, M. (2003, October 9). To Make Sure Votes Count, Sensor Device Goes Back On. New York Times, http://www.nytimes.com/2003/10/09/nyregion/09VOTE.html

Meister, D. (2000). Changing Concepts of Test and Evaluation, Proceedings of the IEA 2000/HFES 2000 Congress (6), p. 554-556

Mercuri R. (2002). Humanizing Voting Interfaces, *Presentation to the UAPA Conference, Orlando, FL*

Mercuri, R. (2000). Electronic Vote Tabulation Checks & Balances. Doctoral dissertation, University of Pennsylvania, Philadelphia, PA

MIL-H-46855, Human Engineering Requirements for Military Systems, Equipment and Facilities

MIL-HDBK-761, Human Engineering Guidelines for Management Information Systems

MIL-STD-1472, Design Criteria Standard Human Engineering

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, p. 81-97

Molich, R. & Jeffries R. (2003). Comparative Expert Reviews *Proceeding of the CHI 2003 Conference,* p. 1060-1061

Mosier, J.N. & Smith, S.L. (1986). Application of Guidelines for Designing User Interface Software *Behaviour and Information Technology Vol. 5, No. 1, p. 39-46.*

ECRI, National Center for Voting Technology (1988). "An Election Administrator's Guide to Computerized Voting Systems," Plymouth Meeting, PA.

Neale, D. C. & Kies, J. K. (2000). *Symposium on Recent Advances in Critical Incidence Techniques, Proceedings of the IEA 2000/HFES 2000 Congress*, p. 6-589.

Nelson, J & Molich, R. (1990). Heurisitic Evaluation of User Interfaces, *Proceedings of ACM CHI'90 Conference on Human Factors in Computing Systems* 1990 p. 249-256

Niemi, R.G. & Herrnson, P.S. (2003). Beyond the Butterfly: The Complexity of U.S. Ballots *Perspectives on Politics*, Vol. 1 p. 317-326.

Olson, G.M. & Moran, T.P. (1998). Commentary on "Damaged Merchandise?", *Human-Computer Interaction*, *Vol. 13, p. 263-323*

Quesenbery, W. (2001). Voting for Usability: A Backgrounder on the Issues, *Talk presented at TECH*COMM 2001 in Washington DC*

Redish, J. (moderator); Bias, R.G. (moderator); Bailey, R.; Molich, R.; Dumas, & J., Spool, J.M. (2002). Usability in Practice: Formative Usability evaluations – Evolution and Revolution, *Proceeding of the CHI 2002 Conference*, p. 885-890

Roth, S.K. (undated). Human Factors Research on Voting Machines and Ballot Designs: An Exploratory Study

Roth, S.K. (1998). Disenfranchised by Design, *Information Design Journal*, Vol. 9, No. 1 p.1-8

Rubin, J.; Miller, J.R.; Wharton, C.; & Uyeda, K.M. (1991). User Interface evaluation in the real world: A Comparison of Four Techniques Proceedings of ACM CHI'91 Conference on Human Factors in Computing Systems, p.119-124

Savage, P. (1996). User Interface evaluation in an Iterative Design Process: A Comparison of Three techniques, (1996) *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems*, v.2 p.307-308

Section 508 Electronic and Information Technology Accessibility Standards (2000). Architectural and Transportation Barriers Compliance Board, 36 CFR Part 1194, http://www.access-board.gov/sec508/508standards.htm

Smith, S. (1986). Standards versus Guidelines for Designing User Interface Software *Behaviour and Information Technology*, Vol. 5, No. 1, p.47-61

Souza, F. & Bevan, N. (1990). The Use of Guidelines in Menu Interface Design: Evaluation of a Draft Standard, *Proceedings of IFIP'90: Human Computer Interaction 1990*, p. 435-440

Spool, J. (2003). (Unpublished) Evolution Trumps Usability Guidelines

Tadayoshi, K.; Stubblefield, A; Rubin, A.D. & Wallach, D.S. (2003). Analysis of an Electronic Voting System

Traugott, M.W. (2002). Testing Alternative Hardware and Ballot Forms, *Prepared for the meeting of the Working Group on Voting Technologies and Balloting*

Voting Irregularities in Florida during the 2000 Presidential Election (2001, June). U.S. Commission on Civil Rights Report, http://www.usccr.gov/pubs/vote2000/report/main.htm

Wald, A. (1947). Sequential Analysis, John Wiley

Wilson, S.V. (2003). Opinion on “Southwest Voter Registration Education, et al vs. Kevin Shelley, in his official capacity as California Secretary of State”, *U.S. 9th District Court of Appeal*

Appendix A – Glossary

The purpose of this glossary is to clarify the terminology used in this report; the definitions are not to be taken as an officially approved general-purpose standard. Moreover, the scope of this glossary is limited to those terms needed in a discussion of voting and usability and the definitions given are to be understood within that context. The glossary does **not** cover other voting areas, such as registration or security.

Accessibility

Accessibility is a measurable characteristic that indicates the degree to which a system is available to, and usable by, individuals with disabilities. The HAVA also includes accessibility for Native American or Alaska Native citizens and alternative language access for voters with limited proficiency in the English language.

Acceptance Testing

The examination of a voting system and its components by the purchasing election authority (usually in a simulated-use environment) to validate performance of delivered units in accordance with procurement requirements, and to validate that the delivered system is, in fact, the certified or qualified system purchased. Testing to validate performance may be less broad than that involved with qualification testing and successful performance for multiple units (precinct count systems) may be inferred from a sample test.

Ballot

A form presenting a sequence of contests.

Ballot Image

An electronically produced record of all votes cast by a single voter.

Candidate

A person contending in a race for office. A candidate may be explicitly presented as one of the choices on the ballot, or may be a write-in candidate.

Certification by ITA

Occurs when an ITA (or other authorized agent) formally asserts that a product is qualified according to established criteria. For voting systems the established criteria are the voting system standards (VSS).

Certification Testing

The state examination, and possibly testing, of a voting system to determine its compliance with state laws, regulations, and rules and any other state requirements for voting systems.

Conformance

The degree to which a product or other object meets the explicit requirements of a standard.

Contest

A decision to be made within an election. May be either a race for office or a referendum. A single ballot may contain one or more contests.

Direct Recording Electronic (DRE) Voting System

A voting system that records votes by means of a ballot display provided with mechanical or electro-optical components that can be actuated by the voter; that processes the data by means of a computer program; and that records voting data and ballot images in internal and/or external memory components. It produces a tabulation of the voting data stored in a removable memory component and in printed copy.

Election

The activities whose purpose is to ascertain on a single occasion the intent of the voters in one or more contests. It includes verifying voters as registrants, allowing them to cast votes, and tallying the results.

EAC

Election Assistance Commission

FEC

Federal Election Commission. Home page at http://www.fec.gov

HAVA

The Help America Vote Act of 2002, Public Law 107-252. Full text at http://fecweb1.fec.gov/hava/hava.htm

Human-Computer Interaction

A discipline concerned with the design, evaluation, and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them. Also, a collection of behaviors and responses that occur between a computer and a human attempting to accomplish a task. For the human (user) this involves both physical and psychological processes.

Human Factors (Ergonomics)

"The scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and methods to design in order to optimize human well-being and overall system performance." (Source: International Ergonomics Association)

Independent Testing Authority (ITA)

A third-party laboratory (neither vendor nor procurer) accredited to assess the conformance of a given product to a standard. For voting products, ITAs verify that the voting product conforms to the voting system standards or VSS.

IEEE

Institute of Electrical and Electronics Engineers. Home page at http://www.ieee.org

ISO

International Organization for Standardization. Home page at http://www.iso.org

Multi-seat Contest

Contest in which multiple candidates can win, up to a specified number of seats. Voters may vote for no more than that specified number of candidates.

NASED

National Association of State Election Directors. Home page at http://www.nased.org

Null Vote

Occurs when none of the alternatives in a given contest is selected.

Null Ballot

Occurs when there is a null vote for every contest on the ballot.

Overvote

Occurs when the number of alternatives selected by a voter in a contest exceeds the maximum number allowed for that contest.

Polling Place

The area within the polling location (physical address) where voters cast ballots.

Qualification Testing

The examination and testing of a computerized voting system by an Independent Test Authority to determine if the system complies with the qualification performance and test standards and with its own specifications. This process occurs prior to state certification.

Referendum

A contest between two (or more) choices in response to a question (e.g. bond issue, recall, retention of a judge in office, proposed amendment).

Residual Vote

The total number of votes that cannot be counted for a specific contest. There may be multiple reasons for residual votes (e.g., overvoting a contest, failure to cast ballot before leaving polling place). .

Rolloff

The difference between number of votes cast for contests in the higher offices on the ballot and the number cast for contests that are lower on the ballot. It is sometimes referred to as voter fatigue.

Self-Contained, Closed Products vs. Open Architectures

We are using the Access Board's definition. Self-contained, closed products are those that generally have embedded software and are commonly designed in such a fashion that a user cannot easily attach or install assistive technology. These products include, but are not limited to, information kiosks and information transaction machines, copiers, printers, calculators, fax machines, and other similar types of products. These are products that are expected to contain all the accessibility features necessary for use by persons with disabilities. This is contrasted with "open architecture" products for which the end user is intended to provide some form of adaptive technology (e.g., a screen reader or external braille display).

Single-seat Contest

Winner–takes-all contest in which voters may vote for only one candidate. The candidate with the highest number of votes wins the contest.

Specification

A document that prescribes technical requirements to be fulfilled by a product, process, or service.

Spoiled Ballot

A ballot on which it appears that the voter attempted to make a selection but did so incorrectly. For example, overvoting a contest or including extra markings on a paper ballot would result in a "spoiled ballot."

Standard

A specification that is formally approved by some standards organization or other authority.

TGDC

Technical Guidelines Development Committee.

Undervote

Occurs when the number of alternatives selected by a voter in a contest is less than maximum number allowed for that contest.

Usability

A measure of the effectiveness, efficiency, and satisfaction achieved by a specified set of users performing specified tasks with a given product.

Usability Engineering

A methodical engineering approach to user interface design and evaluation focusing on usability issues.

Usability Testing

A method by which users of a product are asked to perform certain tasks in an effort to measure the product's usability. Typically, formative (or diagnostic) usability testing is conducted as part of a product development process and summative (or empirical) testing is conducted after a product is completed.

Voting Product

One component of a voting system. In this report, the term is used to refer to a product procured from a vendor, such as a DRE terminal.

Voting System

Combination of environment, equipment, ballot, voters, and other persons (e.g., poll workers and election officials) involved in the voting process.

Voting System Standards (VSS)

The Federal guidelines for voting systems, last revised by the FEC in 2002, freely available from http://www.fec.gov/pages/vssfinal/vss.html. Conformance to the VSS is a prerequisite for certification by some states.

Appendix B – Developing and Conducting Usability Conformance Testing Procedures

We have done some preliminary work on the development of the usability test procedures we believe would be necessary to ensure usability. An outline of this set of test processes is provided below. Additional research is necessary to validate our assumptions and initial conclusions and to make specific detailed recommendations for the tests.

B.1 Test Environment

To properly conduct a usability test, the test environment must recreate, to the extent possible, the total system under test. This means using actual users in realistic environments performing the anticipated tasks. For usability testing of voting equipment, a large scale test conducted in a fully simulated, mock-election would be an ideal approach, but such a test has some drawbacks. It requires a large number of participants in order to ensure that all situations that are likely to be encountered in an actual election are encountered in sufficient numbers to draw conclusions. In addition, it is difficult to know the intent of the participants. Finally, total observation and logging of all the results is difficult without equipment modifications.

To minimize the cost and overcome some limitations associated with mock-election testing, we believe usability testing should be conducted at a "subsystem" level before testing at the system level. For subsystem testing, we believe two usability tests should be conducted: voter subsystem testing and poll worker subsystem testing.

B.2 Voter Subsystem Testing

A voting subsystem test conducted to validate the usability of the interaction between the voting product and the user casting a vote allows for isolation of this process from other factors such as differences in poll worker assistance, voter training on the equipment, prior equipment experience, etc. In addition, participants could be provided with specific tasks to perform intended to exercise all of the product capabilities (e.g., the ability to cast a vote and then change it before casting the ballot) as well as intentional error attempts (attempt to overvote an election). Several tasks could be combined into full scenarios to reduce the number of participants required. And, in addition to the ability to ensure participant intent, participants could be more easily monitored in this setting. Ideally, all interactions with the product are logged and time-stamped allowing very detailed metrics to be derived. A sufficiently large number of participants is still needed for detecting errors for the diverse population, but the ability to specify tasks and isolate the participants from the other variables greatly reduces this number.

B.3 Poll Worker Subsystem Testing

The second type is the poll worker subsystem test, which is conducted to validate the usability of the interaction between the voting product and the poll worker during equipment set-up, break down, and during assistance calls from the voter. Similarly, a sufficient number of representative participants are needed to be included, but this number is considerably smaller than that required in a full mock election.

The subsystem tests are conducted against each of the products independently, though several could be tested in parallel since the test process will be defined to be repeatable, reproducible and objective (i.e., independent of the personnel conducting the test and the location of the test).

B.4 Full System Testing

For the full system testing, a full mock-election test is conducted, but the subsystem tests eliminate the testing of a product with serious, known problems. In addition, it might be possible to organize regular mock elections and reduce the overhead associated with testing multiple products.

B.5 Standard Test Materials

Standard test materials will be required for use in both the subsystem test and the system test. For example, sample ballots of varying complexity need to be developed and would become the standard test materials for the test procedures and would be used for all tests for all voting products. Other materials also have to be developed to ensure the validity and reliability of the test: these include instructions presented prior to testing, equipment training, wording of tasks, presentation order of tasks, assistance provided, etc. Test instructions need to be standardized as would any help material provided during testing. Finally, the recruiting information (the nature and number of participants) to be used needs to be specified and standardized.

B.6 Feasibility and Limitations

Differences are anticipated between the results of the test and the results in a real election, but this is unavoidable. For example, real ballots should be used in testing and we recommend that the sample ballots cover a range of complexity and length. However, the results of testing (the actual numbers of errors expected) may vary from that experienced in an actual election if the real ballot were of exceptional size or complexity or poorly designed. Similarly, we recommend the testing be conducted in a setting closely approximating a realistic polling place in terms of physical layout and ambient conditions (lighting, temperature, and acoustics). However, differences in results may be found if the product is used in a non-standard

environment (e.g., under poor lighting conditions). It is anticipated there would be an added cost for conducting the usability tests in addition to that for conducting the current hardware and software ITA tests. However, there should not be any additional elapsed time associated with the usability tests since they can be conducted in parallel with the present ITA tests, provided sufficient equipment is available. The actual cost of conducting these usability tests is not known at this time, and any research into the development of the test procedures would need to provide cost estimates and cost reduction recommendations. In the short term, any work performed would provide valuable information on the nature and extent of usability problems and the resultant errors that are currently being experienced. Preliminary findings from the testing could be provided to vendors in an open forum (without attributing the results to a specific product); this will assist the vendors in understanding the issues of usability associated with voting and to help them develop better products. Vendors would also be able to adopt small scale versions of the tests as formative or diagnostic tests to be conducted as part of their own internal design process and for internal summative tests prior to qualification.

Appendix C – Statistical Data Analysis

This Appendix addresses the question of how many participants would be needed in a usability test of voting products in order to make reliable estimates of presumably low error rates. As we have argued in the report, a controlled experiment with a valid sample of users is the only reliable way to directly measure bottom-line metrics of system performance, such as error rates and time on task.

In contrast to some previous studies, we will not aim to estimate the mean error rates for specific population groups. There also has been work (suggested specifically in the context of ballot voting problems) to find a "reasonable means of estimating the number of subjects required" for testing. Bailey (2000) uses binomial probability models in a diagnostic testing scheme that seeks to bring in enough subjects to trigger all the existing errors; each system error is presumed to have a fixed probability of being triggered by any individual test subject.

Bailey estimated that if, in the 2000 presidential election, the infamous butterfly ballot caused 1% of votes to be inadvertently cast incorrectly, one could assume that usability testers would each have a 1% percent chance of uncovering that error during testing; conversely, 99% of subjects would not be affected by that problem.

Given that there are n subjects, each with a probability p of encountering the problem, then the probability of that problem being triggered by at least one of the n subjects is q(p,n)= 1-(1-p)^n. Bailey shows how large a sample would be needed to uncover the problem with a certain probability by putting p=.01 and varying n. For example, if n=289 subjects, then the probability of at least one of them uncovering the problem is q(.01,289)= .95. Similarly, setting q(.01,n)=.99 requires n to be at least 423. Bailey's exposition states that the above numbers show that 289 testing subjects would be needed to find 95% of such problems, and 423 subjects are needed to find 99% of the problems. We presume that he supposes that there are n testing subjects, and each problem has an independent discovery rate of 1% by each subject; then each problem is discovered by at least 1 subject with probability q(p,n). In that case, the average number of problems with that discovery rate found would be q(p,n) of the those problems. Of course, the numbers of problems present in a voting system and their respective discovery rates will not be known before the testing occurs.

We propose instead to test a voting system by using a test to determine if the system's failure rate is acceptably low, where the failure rate is the proportion of the population that fails to use the system successfully for any reason. The description of the Wald test that follows shows that if the acceptable error

rates can be pre-determined, it is possible to do sequential testing that can limit the number of subjects that must participate.

Sequential testing, pioneered by Abraham Wald (Wald, 1947), was considered important enough to be classified during World War II, where it was used for sampling inspection of manufactured goods. In certain situations, Wald's Sequential Probability test can save time and money by limiting the number of subjects needed for testing. The testing of a system or a manufactured lot of products can be modeled by sampling from a binomial population with failure rate p (with p between 0 and 1); that is, independent subjects tested have a probability p of failing the test and a probability (1-p) of passing the test. The goal of the testing is to determine whether the failure rate is above or below acceptable limits. In contrast, conventional tests would test a fixed sample of subjects, and the lot or system would pass or fail depending on the results of the entire sample.

In certain cases when the samples are tested in sequence, the results can be such that a firm conclusion can be reached without having the need to run the rest of the subjects. For instance, suppose we test a system to see if its failure rate is below 0.01 and schedule 25 subjects. If the subjects are tested sequentially, and 5 of the first 6 subjects fail the system, then the system will flunk regardless of the result of the next 19 trials, which thus become unnecessary. Sequential testing has been incorporated, though not without controversy, in some clinical tests of new medical procedures, where reducing the number of subjects may well save lives. We suggest that sequential testing may also be applied to testing voting products for usability (and, in fact, this technique is part of the ITA testing for hardware and software compliance).

In Wald's sequential tests, the procedures for reaching a conclusion and stopping the test are not haphazard but spelled out in advance given what failure and error rates are acceptable. At each stage of the test, the number of failures up to that point is tracked and compared to a pre-specified threshold for that stage. If the number of failures is greater than the rejection threshold, then the system is considered to have failed. If the number of failures is smaller than the acceptance threshold, then the system is accepted. If the number of failures is between the thresholds, then the test continues to the next stage, with new thresholds applying to the new stage.

The form and thresholds for the sequential probability test depend on several predetermined parameters, which are listed here with discussion below:

- p0= Highest failure rate (proportion) we are willing to accept; a system with failure rate that is no greater than p0 is acceptable.

- p1= Lowest failure rate (proportion) we find unacceptable; if a system has failure rate p1 or higher, then it should be rejected.

- alpha= (Maximum) Probability of rejecting a (minimally) acceptable system

- beta= (Maximum) Probability of accepting an unacceptable system

The actual results of the test depend on the real failure rate p, which is the proportion of subjects in the tested population that would fail the test. The subjects should be independent of each other and serve as random samples from the population.

We can determine what values of alpha and beta are acceptable, which will include our thinking about the harm created by making each kind of mistake. If many of these kinds of tests are run, both kinds of mistakes are likely to occur occasionally (just as when flipping a fair coin repeatedly, you will sometimes get 5 heads in a row). The chosen values of alpha, beta, p0, and p1 determine the thresholds and stopping points of the tests. Having smaller alpha and beta require longer runs, and indicate less willingness to risk choosing the wrong conclusion. The formulas for the parameter-dependent thresholds can be complicated but are easily computed.

For example, suppose that the maximum acceptable failure rate of a voting system is p0=0.001, but that its real failure rate p happens to be the minimum unacceptable failure rate p1=.01. Suppose also that we want both the false rejection rate alpha and the false acceptance rate beta to be bounded by 0.05. In that case, the average number of subjects needed would be 189; the actual number of subjects needed would vary randomly according to the results. If we wanted both alpha and beta to be 0.01, then the average number of subjects needed would rise to 321. If instead we relaxed both alpha and beta to be 0.1, then the average number of trials would be 125. Relaxing alpha and beta even further to .25 reduces the average number of subjects needed to 40.

In addition to alpha and beta, the choices of p0 and p1, and how they relate to the real error rate p, also affects how many subjects will be needed. If one of p0 or p1 is obviously wrong, then the test can terminate speedily. For instance, if p0 is .001, and half the runs are failures, then the test can terminate quickly. However, for very low p0, it can take many trials without error to convince the test that the real p is less than or equal to p0, especially if p1 is relatively close to p0. In general, increasing the ratio of p1 to p0 will reduce the average needed number of trials. As an example, suppose again that alpha=beta=.05, p1=.01, and p0=.0001 rather than .001. If the real p=p1=.01, the average needed number of subjects is only 74. However, suppose the real failure rate p=0. Then the test takes 296 subjects, because when p0 is tiny, it takes many trials to convince the test that the failure rate is really that small, unless the alternative p1 is so large as to be obviously untenable.

The specific implementation (i.e., choosing the "acceptable" values for p0, p1, alpha, and beta to be used in the Wald process) on the voting product testing with users will need to be determined. One of the purposes of the research proposed above (see Section 6.4) would be to gather data about the actual error rates that could then be used as a guide for determining meaningful and realistic thresholds for conformance tests (see Section 6.10).

Appendix D – Report Methodology

Writing this report required expertise in human factors and ergonomics, usability and accessibility of information technology, voting systems, standards development, conformance testing, and statistics. It also required talking to representative stakeholders from across the election and voting communities in order to identify relevant issues and map these to research and best practices that could be applied to voting systems. NIST created a team that had the necessary expertise and analysis skills in June of 2003. In this Appendix we provide a description of the methodology we used to do the analysis and write this report. Appendix E contains the biographies of the authors.

It was critical to understand the human factor, usability, and accessibility issues from the perspectives of the many different stakeholders in the elections and voting process. The challenge was to then understand the current situation for voting systems and to identify what approaches for general research and best practices could be brought to bear to improve the usability and accessibility of voting systems.

The voting team spent several months reading the relevant literature and talking to numerous individuals knowledgeable about elections and voting systems. We reviewed the research and best practices literature[17] in the following general areas:

- Human factors and usability,
- Accessibility for the disabled,
- User interface standards and guidelines,
- Accessibility standards and guidelines (e.g., Section 508 and the Web Accessibility Initiative),
- Testing and evaluation methods for usability and accessibility, and
- Conformance test processes for standards.

We also reviewed the literature specifically for voting systems and usability and accessibility, much of which exists as research papers, news articles, websites, workshops held since the 2000 elections, vendor demonstrations and literature, and email reflectors/discussion boards on electronic voting (e.g., upa-evoting@yahoogroups.com and verifiedvoting.org). Topics covered can be categorized as:

[17] Note that the references cited in this report are only those that are directly pertinent to the report and are just a subset of the literature that was actually examined.

- Existing voting standards for usability and accessibility and ITA accreditation process,

- Sample ballots and state laws insofar as they affect voter experience with a voting system,

- Vendor voting products,

- Evaluations of usability and accessibility of voting products,

- Research papers on voting and human factors, usability and accessibility, and

- News stories about the voter experience and poor usability.

We talked to representative stakeholders and researchers associated with the election and voting communities. This included visits, discussions and phone calls at NIST and elsewhere involving election officials, vendors, voter advocacy groups, and researchers; attendance at various technical meeting such as the 2003 IACREOT Conference and Trade Show; the 2003 ACM Computer Human Interaction Conference; the August 2003 ACM Voter Verification Workshop; and IEEE standards meetings and teleconferences. We tried to speak with anyone and everyone who had looked at aspects of usability and accessibility for voting systems or, at a minimum, read their writings. For example, we participated in the following activities:

- Meetings with FEC officials (for example, Penelope Bonsall) and other government personnel such as US Access Board staff, and Eric Fischer from the Congressional Research Service.

- Meetings with a number of State election officials including representatives from NASED and NASS at NIST and elsewhere, including Doug Lewis and Tom Wilkey,

- Informal discussions with poll workers,

- Discussions with vendors who visited NIST or who attended other voting and election related meetings,

- Informal evaluations: we took the opportunity at the 2003 IACREOT Trade Show to try out every voting product being demonstrated, and had the opportunity to look at some products in other venues,

- Discussions with representatives from the disabilities community. including Steven Booth at the National Federation of the Blind who

evaluated a number of vendor products, Jim Dickson of the America Association of People with Disabilities' Disabilities Vote Project, and David Baquis of the US Access Board, among others,

- Discussions with researchers who have examined usability and accessibility issues for voting, including Ted Selker from the MIT/CalTech Voting Technology Project, Paul Herrnson and his research team from the Universities of Maryland, Michigan, and Rochester, Gregg Vanderheiden, the director of the TRACE R&D Center at the University of Wisconsin who has developed designs for accessible voting DREs, and other researchers,

- Meetings and teleconferences with members of the IEEE P1583 Accessibility and Usability Task Group, and

- Usability and human factors professionals with an interest in voting systems, such as UPA and HFES members.

We also reviewed the current ITA testing process for certification of voting products and the current vendor system engineering processes for user-centered design, and usability and accessibility testing. We then identified the gaps between industry best practices and research (for both standards development and usability and accessibility design and testing) and the current situation for voting products and systems. By analyzing these gaps, we were then able to define a set of recommendations for improving the usability and accessibility of voting systems.

Appendix E – Author Biographies

Dr. Sharon Laskowski

Sharon Laskowski is the main author as well as editor of this report. She is a computer scientist in the Information Technology Laboratory of the National Institute of Standards and Technology where she manages the Visualization and Usability Group in the Information Access Division. The mission of the Division is to accelerate the development of technologies that allow intuitive, efficient access, manipulation, and exchange of complex information by facilitating the creation of measurement methods and standards. In particular, the Visualization and Usability Group is developing evaluation methods, metrics, and standards for human-computer interaction.

Dr. Laskowski's work on investigating standards and conformance testing issues for usability and accessibility of voting systems included participation on NIST's pre-HAVA, ad hoc voting issues team in 2002, the FEC Advisory Board on Usability and Human Interface Standards, and the IEEE P1583 Usability and Accessibility Task Group. She also organized and moderated the panel on usability and accessibility for the December 2003 NIST Conference on Building Trust and Confidence in Voting Systems.

Other recent work has focused on usability evaluation methods and standards such as the development of ANSI/INCITS Standard 345-2001, the Common Industry Format for Usability Test Reports, which NIST developed with human factors and usability engineering industry leaders as part of the Industry Usability Reporting Project. She has provided advice on a number of accessibility activities related to the Section 508 IT accessibility requirements and the development of the INCITS V2 standard protocol for more transparent accessibility. She created the NIST Web Metrics project for experimenting with rapid, remote, and automated web usability evaluation that includes tools for user logging and category analysis. She has contributed to information visualization research, in particular for large document collections.

Over the years, she has been an active researcher in a number of other areas of computer science including expert systems, plan recognition, analysis of algorithms, and computational complexity. She is a member of the Usability Professionals' Association (UPA), the Institute of Electrical and Electronics Engineers (IEEE), the Association for Computing Machinery's Special Interest Group on Human Computer Interaction (ACM SIGCHI), and a founding member of the local chapter of SIGCHI: DCCHI.

Prior to joining NIST in 1994, Dr. Laskowski was a lead scientist at the MITRE Corporation. She has also been an assistant professor in the Computer Science Department at the Pennsylvania State University. Dr. Laskowski received her BS degree in Mathematics from Trinity College, Hartford, CT and her PhD in Computer Science from Yale University.

Dr. Marguerite Autry

Marguerite Autry, a Senior Human Factors Engineer at User-Centered Design, has a Ph.D. in experimental psychology and seven years of usability experience. She has conducted usability evaluations and non-user based evaluations for a number of commercial and governmental clients. She has worked with other clients such as General Electric, iXF, National Association of Realtors, Centers for Disease Control and Prevention, Food and Drug Administration, Health Resources and Services Administration (with projects for both HIV/AIDS Bureau and Bureau of Health Professions), National Library of Medicine, National Institutes of Health, National Cancer Institute, National Institute on Alcohol Abuse and Alcoholism, National Institute of Arthritis and Musculoskeletal and Skin Diseases, Agency for Healthcare Research and Quality, and Department of Labor. Her scientific background, an undergraduate degree in chemistry and an MS and PhD in experimental psychology, serves clients well in designing and carrying out research experiments and testing.

John Cugini

As NIST prepared to respond to the demand for better voting systems, Mr. Cugini researched the usability issues as a member of NIST's ad hoc voting issues team on voting issues until his retirement from NIST in March of 2003. This work included:

- Participation in the development of a voting model
- Drafting the section on usability for NIST internal voting report
- Representing NIST at several conferences on voting systems
- Representing NIST on the FEC's Advisory Panel on Usability and Human Interface Standards

From 2000 until 2003, his work focused on the interaction between visualization and usability. In particular, he was a major contributor to NIST Web Metrics project (http://www.nist.gov/webmetrics). This work included design and implementation of software that analyzes how users interact with a given website. From 1994 until 2000, he worked on the development and evaluation of prototypes for information visualization, with particular application to document browsing and searching.

From 1988 to 1994, his major effort was the construction of conformance tests for the PHIGS standard. PHIGS is a complex standard describing an application programming interface for 3D graphics. Measuring conformance involves an interactive feedback loop in which a human operator must recognize visual features of the 3D display.

Starting in 1979, his work at NIST was in the area of programming language standards. This included development of test sets for implementations of BASIC and FORTRAN, standardization of numeric accuracy, impact analysis of the revision to COBOL, and a survey publication evaluating several major programming languages. He has participated actively in national and international standards organizations, including those for Ada, BASIC, C, and Common Lisp. From 1984 to 1988, his primary work was research on expert systems. This included evaluation of the KBS-oriented languages, Lisp, Prolog, and OPS5. It also involved a research project that provided conceptual navigation through a knowledge base by means of graphics, using Prolog and GKS.

Mr. Cugini received his AB from Columbia in 1970 with a major in philosophy. He worked for the U.S. Army from 1971 until 1978 as a programmer and instructor. During that time he earned an MS in computer science at the University of Iowa in 1977.

Bill Killam

Bill Killam, MA, CHFP, is the President and Principle Human Factors Engineer at User-Centered Design, Inc. Mr. Killam is board certified in Human Factors Engineering by the Board of Certification in Professional Ergonomics and has been providing Human Factors Engineering, user-centered design, and usability services for over 23 years. He has degrees in both engineering and psychology and has provided product design and testing service to the US Government as well as numerous commercial and non-profit organizations including IBM, GTE, TRW, E-Systems, GEICO, CapitalOne, Nextel, the US Army, the FBI, the Food and Drug Administration, the Center for Disease Control and Prevention, the General Services Administration, the National Cancer Institute, and the Surgeon General. Some of this work has been directed at the Section 508 mandate for IT accessibility, but he has been developing user interfaces for people with disabilities for many years.

He is an active member of the human factors engineering and usability testing community at both the national and local level and has been the Vice President and President of the Potomac Chapter of the Human Factors and Ergonomics Society (HFES), the president of the DC Chapter of the Usability Professionals Association, and is on the board of DC Chapter of the Association of Computing Machinery's Special Interest Group on Human Computer Interaction (ACM SIGCHI). Mr. Killam teaches Human Factors Engineering at both the University of Maryland

and George Mason University and has been a guest lecturer for 4 years at the University of Maryland's HCIL annual open house.

He has authored a number of publications, has been a reviewer for several books on Human Factors, and was a member of the Special Editorial Board for Human Sciences for the British publication *Interacting with Computers*. He was a contributing author for the DOD HCI Style Guide, the DoD's DII Interface Specification, and the author of the DoD AGCCS Style Guide. One of his projects was recently highlighted as a case study in *Interaction Design: Beyond Human-Computer Interaction* (Preece, Rogers, & Sharp, 2002), a new textbook published by John Wiley & Sons.

Dr. James Yen

James Yen is Mathematical Statistician in the Statistical Engineering Division of the Information Technology Laboratory at the National Institute of Standards and Technology in Gaithersburg, Maryland. He obtained his Ph.D. from the Department of Statistics at Stanford University in 1997. He works in the areas of data analysis, applied probability, and computer applications. Dr. Yen consulted with the team on the statistical analysis for the report, primarily for the discussion in Appendix C.