Defining a Summative Usability Test for Voting Systems

A report from the UPA 2004 Workshop on Voting and Usability

September 2004

Report Author:	Whitney Quesenbery
Workshop Participants	Boaz Chen, Louise Ferguson, Bill Killam, Sharon Laskowski, Cheyenna Medina, Richard Niemi Ian Piper, Amy Pogue, Whitney Quesenbery, Janice (Ginny) Redish, Josephine Scott, Mike Traugott, Elizabeth (Dori) Tunstall

This report was submitted as an informational annex to the IEEE P1583 Draft Standard for the Evaluation of Voting Equipment, September 28, 2004.

Usability Professionals' Association

promoting usability concepts and techniques worldwide

140 North Bloomingdale Road, Bloomingdale, IL 60108-1017
Phone 630-980-4997 Web: www.usabilityprofessionals.org

Introduction

At the UPA (Usability Professionals' Association) 2004 annual conference, a group of interdisciplinary professionals held a workshop to consider issues in voting and usability.1 During the workshop, the group tackled the problem of creating a fully-defined usability test protocol for a voting system standard. This paper reports on the approach outlined by that work and the gaps in current knowledge or practice that must be filled to create a fully-defined protocol.

This work builds on the 2003 white paper, "A Proposed Approach to Testing the IEEE Usability/Accessibility Standards," by John O'Hara, which documents methodological issues that must be addressed in designing a usability test for voting systems, and on the Volume 1, Annex C: Usability in the FEC Voting System Standards (VSS) 2002.

Although usability testing is widely employed as part of a user-centered design process, there is little industry experience in usability testing as part of the certification of a system. Such a test must be repeatable, reliable and valid: tests with the same machine should give consistent results, the test protocol should allow for comparison of a system to an objective standard, and the results must be statistically significant, with enough data to reliably uncover even low-incidence errors. In the case of voting systems the tests are used to determine whether users can vote effectively, efficiently, and confidently.

Usability testing can also be used during the pre-election preparations to evaluate the usability of a specific ballot on a voting machine. These usability tests are also important, but would likely use a different methodology[2]. This paper focuses on usability testing used as part of the final evaluation of a voting system to determine whether it meets usability requirements in the standard.

Defining types of usability tests

There are several different types of usability tests, each filling a different role in the design, development, and evaluation of a system. In the white paper, "A Proposed Approach to Testing the IEEE Usability/Accessibility Standards," O'Hara describes two types of usability tests: formative and summative testing.

- Formative usability testing is diagnostic in nature, supporting the development of the design by providing evaluation and feedback on various aspects of the design. This style of testing is typically used at several stages in the design process to help the design team uncover usability problems and make decision about design features. Formative testing is recommended as a best practice for system design in the FEC publications, "Developing a User-Centered Voting System" and "Usability Testing of Voting Systems."

- Summative usability testing is used to evaluate how well the design meets usability requirements. These tests are typically based on the three elements in the definition of usability found in ISO9241-11: effectiveness (completeness and accuracy), efficiency (resources expended), and satisfaction. Summative usability tests usually include larger numbers of participants, allowing for statistical analysis of the performance of the system.

[1] The participants in the UPA2004 Workshop on Voting and Usability were: Boaz Chen, Louise Ferguson, Bill Killam, Sharon Laskowski, Cheyenna Medina, Richard Niemi, Ian Piper, Amy Pogue, Whitney Quesenbery, Janice (Ginny) Redish, Josephine Scott, Mike Traugott, Elizabeth (Dori) Tunstall.

[2] A usability test to evaluate a specific ballot as part of pre-election preparations might be a scaled-down version of this test method, or might use a different approach. This is also a gap that must be filled.

The NIST human factors report adds a third type of testing for consideration: conformance testing

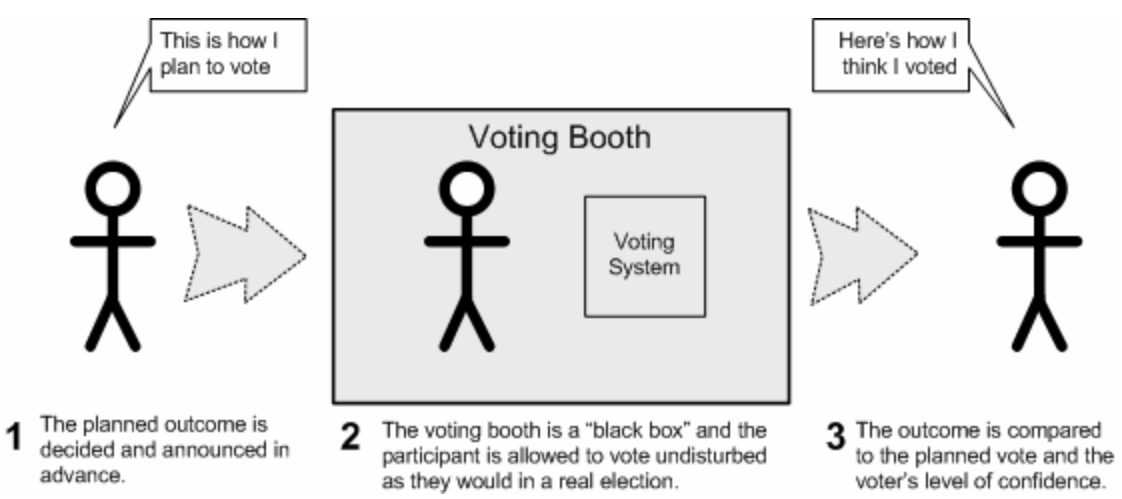
- Conformance tests have the goal of detecting aspects of the system that do not meet a standard, not to identify the cause of failure. A usability test for conformance to a usability standards is a specialized type of summative test. For voting systems, such a test requires usability performance benchmarks as part of the standard to test against, a representative set of test ballots and a well-specified test protocol that is explicit about how many and what types of voter populations to test against. Further, the conformance test has to be validated to produce repeatable results.

Although the workshop participants all endorsed a user-centered design process and the use of formative testing as a method to produce highly usable voting systems, the focus of this work is on the usability evaluation of a completed product. The workshop did not distinguish between the protocol for a summative usability test and a conformance test.

Most importantly, this work attempted to identify the metrics that would be needed to identify a "passing grade" on such a test. Many of these metrics require new research to establish the pass/fail criteria before a full conformance test could be designed.[3] They are identified in "Gap Notes" throughout this paper.

Approach

The goal of this summative usability test for voting systems is to identify and measure failures, not to diagnose their cause. The focus is on the accuracy, efficiency, and confidence of the voting experience. The approach outlined in this paper is based on a testing concept developed by John Cugini of NIST and presented at several industry meetings and symposia.



The structure of the test outline is based on industry best practices. In addition, two standards (one an early draft) were considered:

- ANSI NCITS 354-2001 *Common Industry Format for Usability Test Reports*

[3] Filling these gaps and providing metrics for evaluating the usability of a voting system would be appropriate work for NIST, under its HAVA mandate.

- ISO DTS 20282-2: *Ease of operations of everyday products – Part 2: test methods (for walk up and use and consumer products when used for the first time)*

Defining a Summative Usability Test

For any usability test, the following elements must be defined in the protocol and included in the report of the test results.

1. The purpose of the test
2. The voting system to be used for the test
3. The context of use
4. The characteristics and numbers of test participants
5. The tasks or activities for test participants
6. Data collected or measured during the test
7. How the test data will be analyzed and how a system will pass or fail the test

1. Purpose of the test

The goal of the usability test is to evaluate a single voting system against a set of usability requirements.

This test is not intended for a comparative evaluation of different systems, to provide design recommendations (though any usability test can provide insights into possible improvements in a system), or to establish requirements. Each of these are valid and important types of tests, but would be conducted differently, using a different test protocol.

The September 2004 draft of the IEEE Voting Systems Standard, Section 5.3 identifies the usability goals for a voting system as one that allows voters to cast their vote:

- Correctly – use the voting system to register their intended selections with minimal errors and easily detect and correct errors when they occur.
- Efficiently – complete the voting process in a timely manner and without unproductive, unwanted, incomprehensible, or frustrating interactions with the system.
- Confidently – confident (1) in what actions they had to perform in order to vote, (2) that their votes were correctly recorded by the system and will be correctly counted, and (3) that their privacy is assured.

Gap Note 1:

No absolute requirements or benchmarks for correctness, efficiency or confidence for voting systems currently exist. For example, we know in broad outline that it must be possible to vote in a "timely manner," we do not have a specific definition of the maximum allowable time.

In addition, there has not been enough usability research done on voting systems to identify any specific usability issues that might be different here than on other touchscreen interfaces.

2. Definition of the voting system to be used for the test

For the purposes of this test protocol, the following definitions will be used:

- A ballot consists of the specification of races, issues and voting rules
- The layout is the arrangement of the ballot for presentation to the voter
- The voting device is the machine or paper system used to mark the ballot

For the purposes of a usability test, therefore, a voting system consists of a ballot with a layout used by the voter with a specific voting device. As one participant stated, "The link between the ballot and the device is inextricable" in testing a voting system.

The ballot should be built following the manufacturer's recommended process for the system being tested. When there are options in the layout, the ballot should be designed to best practices for election management. The test report should include a description of these decisions and indicate which ballot design guidelines were used.

Gap Note 2:

There is currently no specification for the contents of a test ballot. These specifications need to be developed and should include several different ballot contents ranging from the simple to the complex, based on the number of races, the inclusion of ballot questions, and the complexity of the voting rules (such as straight-party voting, multiple candidate selection, write-in features, as well as less-used features such as judicial retention or split party voting.) These specifications should reflect a realistic range of elections.

3. Context of use

Most usability tests are conducted in an artificial environment, though it may simulate aspects of the normal context of use. A polling place is typically set up for an election in a space normally used for other purposes, such as a school gymnasium, a firehouse, a room in a community center, a church or civic center. The usability test should be conducted in a similar environment, following best practices for polling places for layout, to ensure privacy and accessibility. Creating a realistic test environment is an important consideration in designing the test.

Standard usability testing procedures for obtaining consent forms and informing participants that their actions will be recorded must be incorporated into the test plan, but this "housekeeping" can be done outside of the mock polling place.

Gap Note 3:

Regulations and guidelines for polling places are established by local jurisdictions. Specifications for the usability test location and setup should be based on the best practices in these guidelines to reflect a realistic range of elections across the country.

The physical environment should be similar for all tests. This includes general environmental factors such as:

- Lighting
- Temperature

- Noise level

As well as considerations such as:

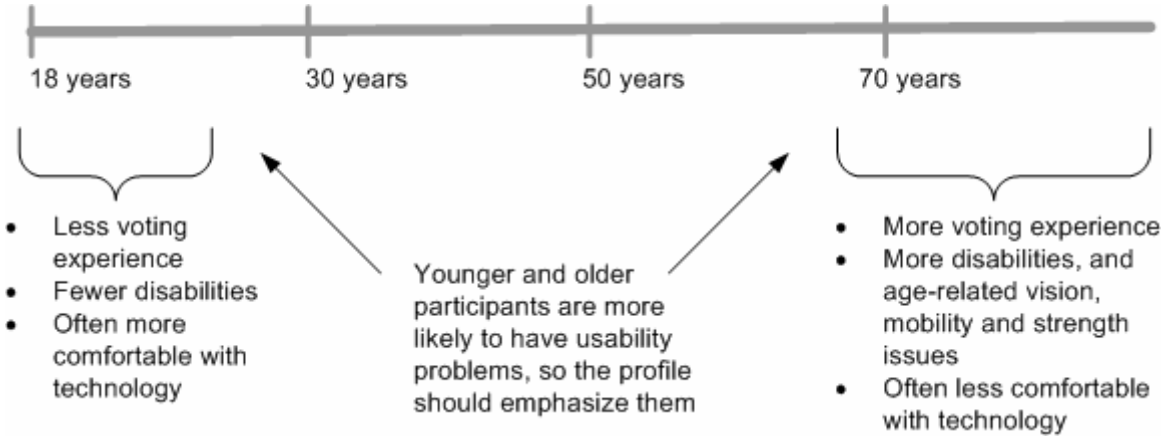
- General room layout
- Proximity of individual voting systems
- Traffic “corridors” around the voting systems
- Number of people in the simulated polling place

4. The characteristics of test participants

The recruiting profile for usability test participants is always a critical factor in conducting a usability test, but is particularly difficult for testing voting systems. In a typical industry usability test, the participant recruiting profile is usually skewed towards the central 80% of the population that might use the product.

For usability test of a voting system, with its comparatively simple task and relatively low error rates, this approach will not work. It is important to ensure that the people at most risk of failure in using a voting system are included in the test. (This includes people with a variety of disabilities, socio-economic backgrounds, language skills and ages.) The solution is to recruit more participants from higher-risk profiles, rather than an even number from each group.

For example, looking at the single variable of age, we might recruit more young and older people than those in their middle years.



Prior voting experience is an important consideration. This includes the number of years as a voter and the type or variety of voting systems the participant has used. In addition, their voter type, that is, their social relationship to elections must be considered. Design for Democracy has identified 5 voter types that might be considered in ensuring that the full spectrum of elections experience is included.

- avid voter
- civic voter
- issue voter

- excluded vote
- apathetic voter

Demographic characteristics to be considered in creating the participant profile include:

- age
- gender
- race/ethnicity
- access or experience with computers and technology
- socio-economic status
- level of education
- rural/urban environment
- region of the country[4]
- primary language
- low literacy / English proficiency

Finally, the participant profile must include a range of disabilities. In considering accessibility, we distinguish between ordinary diversity, those with special needs, and those with extraordinary needs. There is work to do in determining how to identify people with different disabilities and in understanding how they are covered in the participant profiles.

HAVA regulations require at least one accessible voting system at each polling place, so we must also distinguish between general voting devices, which must meet the needs of both ordinary diversity and a range of special needs, and machines intended to meet specific types of accessibility requirements, such as audio ballots. One variable in specifying the participant profile is the types of accessibility options the device offers and whether it is intended to be fully accessible.

Gap Note 4:

There needs to be a standard specification for participant recruiting that combines the variables listed above into a voter segmentation with a moderate number of segments, while still covering the full range of human backgrounds and abilities. This includes the work of grouping types of disabilities into segments that face similar problems in accessing a voting system, and where the disability affects the usability of the system for that participant. This recruiting specification will also define the number of different user segments for the test.

Number of participants

The final question about the characteristics of the participants is the number of people needed to reach statistically significant numbers.

Standard research guidelines suggest 50-100 participants per voter segment. If the test includes more than voting system – a ballot, layout and machine - (for example, testing both a simple and a

[4] There are differences in state election regulations, but also political implications in where the test participants are located to consider.

complex election), the number of participants is calculated by multiplying the number of segments times the number of voting systems. (The group concluded that a single participant should not test more than one voting system, even with a randomized order.)

The final consideration in the number of participants required is the confidence interval desired. This point was mentioned, but not explored further.

To test one voting system:

> # user segments X # ballots X (50-100) = number of participants
>
> e.g.
>
> 5 segments X 2 ballots X (50-100) = 500-1000 participants

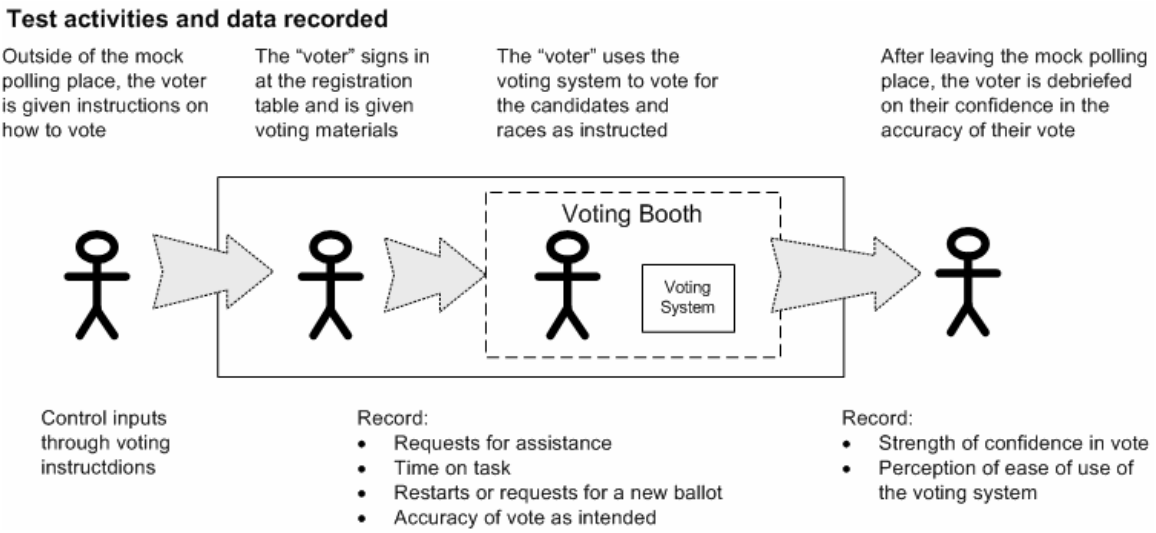
This assumes that each participant completes a single activity: casting a vote on a complete ballot. In fact, each ballot is likely to consist of several races or ballot questions, so each produces multiple data points that can be considered in evaluating accuracy.

This may seem like a large number, but is not beyond the number of voters handled by a single precinct in an election. There is, however, significant work to be done in determining how many (or how few) different types of ballots must be tested or the number of participants required will multiply rapidly.

5. Tasks or activities for test participants

The basic task of a usability test of a voting system is to cast a vote. The test activities are very simple and simulate a typical voting process as closely as possible. In a mock polling place, each participant in the test goes through a complete voting process, including:

- identifying themselves and signing the elections register
- receiving their ballot paper or any other materials required to vote
- receiving any instructions that are part of the normal voting process
- voting as instructed



To ensure that the full functionality of the machine is exercised during the test and to control for the expected results, the participants should be given specific instructions about how to vote, including the candidates or answer to ballot questions. (This is similar to a logic and accuracy test sample set). However, whether the choice of how to vote is left to the participant or dictated to them, the selection must be done in a format that does not directly copy the ballot layout format, for example, using a voter information brochure or other material.[5]

Gap Note 5:

The workshop participants agreed that there would be more control over the test analysis if the participants were given instructions as to how to vote. However we do not know what impact this decision might have on the test outcome. (For example, will participants "work harder" to vote for a candidate they selected (or one for whom they have strong partisan feelings), or one that was selected for them?). This question requires more research for a confident decision.

The group did not address the question of how to select candidate names for the usability test, or how much the artificial choice (because the test is not set in the context of a real election campaign) affects voter performance.

Interaction with test participants

During the test, no attempt is made to interview participants to understand their cognitive or emotional interaction with the voting system. While the voters are in the mock polling place, especially the voting booth, they are not interviewed, given *test* instructions, or prompted to "talk aloud." This is especially important to the measurement of efficiency, especially as the actual task of voting is relatively short. Any observation should be done remotely (for example, through a discrete camera), rather than with a second person in the voting booth with the participant.

Any test instructions, questionnaires or debriefing interviews take place outside of the mock polling place (much as news reporters, campaigners, and others are kept outside of a 50' radius from most polls).

In order for the usability test to be repeatable, the activities and scripts must be detailed and precise. Although it might seem artificial to read from a script, it is important that every participant receive exactly the same instructions (outside of any differences in the details of their task assignments). This includes:

- any pre-test information or general instructions,
- the format, delivery methods and wording of task assignments,
- any instructions given for how to complete the usability test,
- any instructions, training, or practice on how to vote that are given to all participants,
- availability of sample ballots,

[5] It is also possible that the usability test might produce additional functional test data by using the test participants to create randomized paths in an attempt to find functional failures. For example, if a requirement of the system is that it prevent overvotes, participants whose instructions attempt an overvote will exercise this feature. Any failure of a functional requirement during a usability test should be noted and included in the test report.

- any additional instruction, training, or practice offered on the voting system before the participant begins to use the system to vote, and
- any assistance or additional instructions available during the test, with the same answers, instruction or help given for each question that the test participants may ask.

Task complexity

This test uses a single voting device, but might include a range of ballots, ranging from a very simple (few races, with a single selection in each race, first-past-the-post rules) to more complex ballots (including straight party voting, multiple selections in a race, a mix of partisan and non-partisan races, ballot issues, and schemes such as preferential voting).

The use of more than one ballot increases the number of participants required, if (as the group concluded), each participant should use only one voting system (defined as the ballot plus the layout on a device), even if the order of presentation is randomized.

The group discussed options for managing multiple voting systems, including testing the device with a range of ballots, starting with the simplest and moving to the more complex.

6. Data collected or measured during the test

There are three categories of data collected during the test, corresponding to the three broad usability requirements: correctness, efficiency, and confidence.

Correctness

Unlike a real election (where the vote is secret), the test is set up so that the voter can be discretely observed, and their actions inside the voting booth recorded. This allows the test observers to collect information about the voters actions and to compare the vote as cast to the intended vote. This ability to observeis relevant to assessing whether voter confidence in the system is warranted.

There are several different outcomes of the voting activity for each participant.

0. The participant voted as intended.
0. A vote was cast, but there was an error of some kind:
    - The participant did not vote as intended.
    - The ballot was invalid in some way: unintentionally undervoted, overvoted, a spoiled write-in, or some other voter error
0. The participant did not succeed in casting a vote.

The usability test report should include an enumeration of each outcome, as well as the specific types of failures when a vote is cast.

In addition significant actions that indicate possible usability problems are recorded. These might include:

- the number of requests for assistance during voting
- the number of times the voter self-corrects or restarts the ballot.
- the number of times the voter requests a new ballot or requires official assistance to restart a ballot.

Efficiency

Efficiency is measured as a total time on task, either from the time the participant reaches the registration table, or from when they enter the voting booth (or both), until they leave the voting booth. Time waiting to enter the voting booth should not be included, although it may be a part of any real election as we are testing the performance of the voting product not the process in the polling place.

This total time-on-task data includes any time spent reading instructions, receiving assistance, or restarting the ballot, so although these events are counted as part of the accuracy data, their biggest impact on the evaluation of the voting system will be in efficiency.

Confidence

After they have completed voting, participants might be given a standard questionnaire to collect the strength of their confidence in their vote and their overall satisfaction with the voting experience.

Gap Note 6:

Although there are some standard instruments for measuring user satisfaction, none of them are completely suitable for testing voting systems. A standard set of questions (and analysis techniques) needs to be created for this usability test.

7. Specify how the test data will be interpreted

This is the area of this test protocol that is the most difficult. It is possible to report data, and yet have no way to decide whether the results are acceptable (passing) or a failure. Without definitive, quantitative usability measures, the interpretation of the test remains a qualitative analysis, no matter how many quantitative data points are considered.

Gap Note 7:

A method of "scoring" the test, including a method of analyzing the data, must be developed before the usability test can be used for conformance testing. This method must, itself, be tested to establish its validity.

The group came up with several concepts for future consideration.

Test-to-failure

One way to test any product is to test with increasingly difficult or severe conditions until it fails. To determine their failure point, electrical products, for example, are stressed until they fail. This approach could be applied to usability testing voting systems. The voting device would be tested with increasingly complex ballots until the system fails to meet minimum metrics for accuracy, efficiency, or confidence.

The advantage of this approach would be in producing a "confidence zone" -- the range of ballots for which elections officials could be confident in using the machine without additional usability

testing. If this "zone" covered the majority of elections, it would increase overall confidence in the system. More importantly, it would provide officials with a trigger to warn them of when they must take extra care.[6]

Success rates

In testing consumer or commercial products, manufacturers might set an 80% success rate or look for success in the middle 80% of their possible user population. The standards for a voting system must be higher: an entire voting population must be able to use a voting system effectively. In addition, the confidence interval (the likelihood that another test with a different sample of participants will produce the same results) must also be high. No system will have a 100% success rate, so there must be some way of determining how effective is "effective enough" for a voting system.

Creating a "score"

Several different elements must be combined in evaluating the results of this usability test, including time on task, task success, and confidence. Task success, however, can be more than a binary evaluation. This protocol allows for success in the voting task following a usability failure that causes the participant to ask for assistance or re-start the ballot.

The report (and score) might look at the numbers of:

- successes following failure
- partial failures (a vote is cast, but is inaccurate in some way)
- failures (a vote is not cast at all)
- requests (and types of request) for assistance

These values might be weighted for the significance of each failure type in an election. They must also be combined with the efficiency numbers for time on task, and confidence.

One concept[7] is to create a weighted score that takes into account not only the test data, but also the fact that the participant profile is skewed towards groups with a greater risk for usability problems. This score would not be a "percentage of failures" but a more abstract value. This approach would address a political reality: it will never be acceptable to release a report that suggests that a percentage of the voting population will not be able to use a voting system.

Use of the test report as a diagnostic tool

Although the goal of the test is to produce a summative view of the usability of the voting system, the details of the observations might also be useful for both their diagnostic value for the manufacturer and their research value in understanding the usability of voting systems.

Specifically, the types of failures would be valuable if they can be combined with information about the ballot layout, and other details of the voter interaction.

Using the CIF for reporting

ANSI NCITS 354-2001, *Common Industry Format for Usability Test Reports* (also called the CIF) provides a standard template for reporting summative usability tests. The goal of this standard is to allow people reviewing usability tests to compare them easily and it was developed based on industry best practices for summative testing. The group recommends that this format be used for reporting. However, it should be noted that the recommendations in the CIF for number of test

[6] For example, in the case of the butterfly ballot, the confidence zone might have included only the number of candidates in a single race that would fit on one page. The need to use facing pages would have triggered additional testing that to see if the selection alignment was usable.

[7] Boaz Chen provided this concept in group discussions.

participants is not suitable for conformance testing and the CIF does not address how to choose appropriate demographics sets.

Conclusion

Although there are many gaps in the metrics and analysis methods for a summative usability test of a voting system, this paper presents a general approach to creating the protocol for these tests.

This approach is distinguished from formative usability tests and even some summative tests in the following ways:

- It focuses on an overall evaluation of the voting system, identifying usability failures, rather than diagnosing the cause of those failures.
- No interaction with the participants, except that available during a real election, is allowed during the test task.
- The participant profile is designed to be as inclusive as possible, covering the full range of human abilities. However, because of the relatively low error rates, it is skewed towards those likely to have more usability problems in voting.

Related Standards

ISO 9241:1999 - *Ergonomic requirements for office work with visual display terminals.*

This standard provides a definition of usability that is frequently cited in industry publications and other standards:

> Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

ANSI NCITS 354-2001 - *Common Industry Format for Usability Test Reports*

This standard specifies a format for reporting summative usability tests. It was developed by the Industry USability Reporting Project (IUSR) at NIST, and is currently on track to become an ISO standard. Its goal is to "facilitate incorporation of usability as part of the procurement decision-making process for interactive software products. The purpose of the report is to ensure that the product meets usability goals. Examples of decisions include purchasing, upgrading and automating. It provides a common format for human factors engineers and usability professionals in supplier companies to report the methods and results of usability tests to customer organizations." The report outline includes the following sections:

0 Introduction
  0.2 Full Product Description
  0.2 Test Objectives
0 Method
  0.2 Participants
  0.2 Context of Product Use in the Test
    0.2.0 Tasks
    0.2.0 Test Facility
    0.2.0 Participant's Computing Environment
    0.2.0 Test Administrator Tools
  0.2 Experimental Design
    0.2.0 Procedure
    0.2.0 Participant General Instructions
    0.2.0 Participant Task Instructions
  0.2 Usability Metrics
0 Results
  0.2 Data Analysis
    0.2.0 Data Scoring
    0.2.0 Data Reduction
    0.2.0 Statistical Analysis
  0.2 Presentation of the Results
    0.2.0 Performance Results
    0.2.0 Satisfaction Results

References

ANSI (2001). *Common Industry Format for Usability Test Reports* (ANSI NCITS 354-2001). New York: American National Standards Institute, Inc. See also, http://www.nist.gov/iusr/

Department of Justice (2004). *Americans with Disabilities Act: ADA Checklist for Polling Places.* U.S. Department of Justice, Civil Rights Division http://www.usdoj.gov/crt/ada/votingck.htm

FEC (2002). *Usability* (Volume 1, Appendix C of the National Voting System Standard). Washington, D.C.: U.S. Federal Election Commission.

FEC (2003). *Developing a User-Centered Voting System*. Washington, D.C.: U.S. Federal Election Commission.

FEC (2003). *Usability Testing of Voting Systems*. Washington, D.C.: U.S. Federal Election Commission.

IEEE (2004) *Draft Standard on Voting Equipment Design (P-1583).* New York, New York: Institute of Electrical and Electronics Engineers

IEEE (2004). *Draft Section 5.3 Usability and Accessibility Standards* (Draft IEEE Voting Equipment Standard P-1583). New York, New York: Institute of Electrical and Electronics Engineers.

ISO (1997). *Ergonomic requirements for office work with visual display terminals.* (ISO 9241:1999). Geneva, Switzerland: International Organization for Standards. (Parts 10-17 – software interfaces; Part 11 – Guidance on usability; parts 3-9 hardware.)

ISO (1999). *Human-Centered Design Process for Interactive Systems* (ISO 13407:1999). Geneva, Switzerland: International Organization for Standards.

ISO (2004). *Ease of operations of everyday products – Part 2: test methods (for walk up and use and consumer products when used for the first time).* (ISO DTS 20282-2:). Geneva, Switzerland: International Organization for Standards.

Laskowski (2004). *Improving the Usability and Accessibility of Voting Systems and Products.* (NIST Special Publication SP 500-256). Gaithersburg, Maryland: National Institute of Standards and Technology. Also http://vote.nist.gov/Final%20Human%20Factors%20Report%20%205-04.pdf

O'Hara (2003). *A proposed approach to testing the IEEE usability/accessibility standards.* New York, New York: Institute of Electrical and Electronics Engineers.

Public Law 107–252 (Oct. 29, 2002). *Help America Vote Act of 2002 (HAVA)*