Human Factors and Privacy: Report on Final Draft of VVSG

Dr. Sharon Laskowski

Overview

- Summary of significant changes from VVSG 2005
- Review of HFP changes from the previous VVSG Draft
- Usability performance benchmarks

Significant HFP changes from the VVSG 05

- Performance benchmarks (3.2.1.1)
- Poll worker usability (3.2.8)
- Plain language guidance, cognitive requirements (3.2.4-C)
- Accessibility of paper-based vote verification (3.3.1-F)
- Accessibility throughout the voting session (3.3.1-A)
- General adjustability throughout voting session (font size, color, contrast, audio volume, or rate of speech) (3.2.5-E,I, 3.2.3-B, 3.3.3-C.5,8)

Significant HFP changes from the VVSG 05 (Continued)

- Timing requirements (3.2.6.1)
- Low vision more fully addressed and moved to general usability section
  - Require availability of choice of font size and contrast on all VEBD-V machines, not just the accessible-VS. (3.2.5-E)
  - Paper legibility (3.2.5-G)
- The safety requirement now refers to UL 60950-1 (3.2.8.2)
- Usability of the VVSG document improved

Technical Guidelines Development Committee
17 August 2007 Plenary Meeting

There are 14 significant changes since the May Plenary.

- **Updated terminology throughout Section 3.**
  - Including “contest”, “contest choice”, “manufacturer”, “vote”
- **3.1.2: Added definition of “summative usability testing”**
- **3.1.3: Clarified interaction of 3.2 and 3.3**
  - All VEBD requirements apply to the Acc-VS.
- **3.2.1: New metrics and proposed benchmarks**
- **3.2.2-D: Notification of successful ballot casting**
  - “If and only if the voter successfully casts the ballot, then the system ***SHALL*** so notify the voter.”

Changes since the May Plenary (Continued)

- **3.2.2.1-F, 3.2.2.2-F: Ballot casting failure notification**
  - "If the voter takes the appropriate action to cast a ballot, but the system does not accept and record it successfully, including failure to store the ballot image, then the DRE ***SHALL*** so notify the voter and provide clear instruction as to the steps the voter should take to cast the ballot."
  - "If the voter takes the appropriate action to cast a ballot, but the system does not accept and record it successfully, including failure to read the ballot or to transport it into the ballot box, the PCOS ***SHALL*** so notify the voter."
- **3.2.3.1-A: Added discussion of system support of privacy**
  - "The voting system ***SHALL*** prevent others from determining the contents of a ballot."

  Discussion: The voting system itself provides no means by which others can "determine" how one has voted. Of course voters could simply tell someone else for whom they voted, but the system provides no evidence for such statements, and therefore voters cannot be coerced into providing such evidence."

Changes since the May Plenary (Continued)

- **3.2.4-D: Upgraded initial clause in "no bias" requirement**
  - No bias among choices: "Consistent with election law, the voting system ***SHALL*** support a process that does not introduce bias for or against any of the contest choices to be presented to the voter. In both visual and aural formats, the choices ***SHALL*** be presented in an equivalent manner."

- **3.2.5: Clarified "poor reading vision"**
  - "The requirements of this section are designed to minimize perceptual difficulties for the voter. Some of these requirements are designed to assist voters with poor reading vision. These are voters who might have some difficulty in reading normal text, but are not typically classified as having a visual disability and thus might not be inclined to use the accessible voting station."

Changes since the May Plenary (Continued)

- **3.2.5-G, G.1, G.2: Legibility of paper**
  - Upgraded to "shall"
  - Specified 2 "sufficient techniques": font size and magnification

- **3.2.6.1-E, -F: Clarified Voter Inactivity Time and Alert Time**

- **3.2.7-A.3: New requirement in Alternative Language section**
  - "Auditability of records for English readers: Any records, including paper ballots and paper verification records, ***SHALL*** have sufficient information to support auditing by poll workers and others who can read only English."

Changes since the May Plenary (Concluded)

- **3.2.8.2: Maintenance section deleted, in favor of 6.4.5**
- **3.2.8.2-A: Updated citation to UL safety standard**
- **3.3.2: Using "low vision" rather than "partial vision" to conform to common practice**

Usability Performance Requirements

- Goal: To develop a test method to distinguish systems with poor usability from those with good usability
  - Based on performance not evaluation of the design
  - Reliably detects all the types of errors one might see when voters **interact** with a voting system
  - Reproducible by test laboratories
  - Technology-independent
- Given such a test method, benchmarks can be calculated: a system meeting the benchmarks has good usability and passes the test
  - The values chosen for the benchmarks become the performance requirements

Usability testing for certification in a lab

- We are measuring the **performance of the system** in a **lab** so we must control for other variables, including the test participants
- The test ballot is designed to detect different types of usability errors and be typical of many types of ballots
- The test is done in a lab and the environment is tightly controlled, e.g., for lighting, setup, instructions, no assistance
- The test participants are chosen to reliably detect the same performance on the same system
- Test participants are told exactly how to vote, so errors can be measured
- The test results measure relative degree of usability between systems and are NOT intended to predict performance in a specific election
  - Ballot is different
  - Environment is different (e.g, help is provided)
  - Voter demographics are different
- A general sample of the US voting population is never truly representative because all elections are “local”.

Components of the Test Method (Voting Performance Protocol)

- Well-defined test protocol that describes the number and characteristics of the “voters” participating in the test and how to conduct test,
- Test ballot that is relatively complex to ensure the entire voting system is evaluated and significant errors detected,
- Instructions to the “voters” on exactly how to vote so that errors can be accurately counted,
- Description of the test environment,
- Method of analyzing and reporting the results, and
- Performance benchmarks and their associated threshold values.

Performance Benchmarks: Recap of Research

- Validity: tested on 2 different systems with 47 participants
  - Test protocol detected differences between systems, produces errors that were expected.
- Repeatability/Reliability: 4 tests on same system, 195 participants, similar results
- Demographics
  - Eligible to vote in the US
  - Gender: 60% female , 40% male
  - Race: 20 % African American, 70% Caucasian, 10% Hispanic
  - Education: 20 % some college, 50% college graduate, 30% post graduate
  - Age: 30% 25-34 yrs., 35% 35-44 yrs., 35 % 45-54 yrs.
  - Geographic Distribution: 80% VA, 10% MD, 10% DC

Benchmark Tests

- 4 systems, May 19-20, June 1-2
  - Selection of DREs, EBMs, PCOS
- 187 test participants
- 5 measurements
  - 3 benchmark thresholds
  - 2 values to be reported only

The Performance Measures

Base Accuracy Score

- We first count the number of errors test participants made on the test ballot – there are 28 voting opportunities: count how many were correct for each participant
- We then calculate a **Base Accuracy Score:** the mean percentage of all ballot choices that are correctly cast by the test participants

We calculate 3 effectiveness measures

Total Completion Score

- The percentage of test participants who were able to complete the process of voting and having their ballot choices recorded by the system.

Voter Inclusion Index (VII)*

- A measure of overall voting accuracy that uses the Base Accuracy Score and the standard deviation.
  - If 2 systems have the same Base Accuracy Score (BAS), the system with the larger variability gets a lower VII.
  - The formula, where S is the standard deviation and LSL is a lower specification limit to spread out the measurement (we used .85), is:

$$VII = \frac{BAS - LSL}{3S}$$

*range is 0 to ~1, assuming best value is 100% BAS, S=.05, but may be higher

Perfect Ballot Index (PBI)*

- The ratio of the number of cast ballots containing no erroneous votes to the number of cast ballots containing at least one error.
  - This measure deliberately magnifies the effect of even a single error. It identifies those systems that may have a high Base Accuracy Score, but still have at least one error made by many participants.
  - This might be caused by a single voting system design problem, causing a similar error by the participants. The higher the value of the index, the better the performance of the system.

*range is 0 to infinity, if no errors at all.

Efficiency and Confidence Measures

- **Average Voting Session Time** – mean time taken for test participants to complete the process of activating, filling out, and casting the ballot.
- **Average Voter Confidence** – mean confidence level expressed by the voters that they believed they voted correctly and the system successfully recorded their votes.
- Neither of these measures were correlated with effectiveness.
- Most people were confident in the system and their ability to use the system.

Benchmark test results

	Number of Participants Completing The Ballot	Total Completion Score (%) Confidence Intervals (95 % level)	Base Accuracy Score (%) Mean, Standard Deviation	Voter Inclusion Index With 85% LSL Confidence Intervals (95 % level)
System A	50 of 52 (96.2%)	86.3-99.7	95.0, 11	.19-.41
System B	42 of 42 (100%)	92.8-100	96.0, 6	.49-.85
System C	43 of 43 (100%)	92.9-100	92.4, 13	.08-.30
System D	47 of 50 (94.0%)	83.2-98.6	92.4, 19	.03-.22

Benchmark test results

	Number of Participants with Perfect Ballot Including Percent and Index using Adjusted Wald Method	Perfect Ballot Index Confidence Intervals (95 % level)	Voting Time (secs) Mean, Standard Deviation	Participant Confidence (1-5) Mean, Standard Deviation
System A	29 of 50 (58.0%) Index: 1.35	0.79 – 2.40	638.1, 166.1	4.0, 1.0
System B	24 of 42 (57.1%) Index: 1.30	0.73 – 2.44	429.3, 156.3	3.3, 1.4
System C	15 of 43 (34.9%) Index: 0.57	0.29 – 1.00	870.7, 236.0	3.6, 1.4
System D	31 of 47(66%) Index: 1.84	1.07 – 3.52	744.7, 209.3	3.8, 1.2

Proposed benchmark thresholds

- Voting systems, when tested by laboratories designated by the EAC using the methodology specified in this paper, must meet or exceed ALL these benchmarks:
  - Total Completion Score of 98%
  - Voter Inclusion Index of .35
  - Perfect Ballot Index of 2.33

- Systems C and D fail.

- Report time and confidence

- Draft VVSG has placeholders for the values, e.g.,
  "3.2.1.1-A Total completion performance: The system ***SHALL*** achieve a total completion score of at least XXX% as measured by the VPP."

How "tough" should the benchmark thresholds be?

- The benchmark data here used 50 test participants, but the test protocol will call for 100 (to allow statistical assumption of normal distribution to calculate the VII confidence intervals)
  - 100 participants will narrow the confidence intervals and thereby toughen the test.

- **Two points of view:**
  - Proposed benchmarks do weed out poorly performing systems (and, it is relatively easy to raise thresholds)

  vs.
  - This should be a forward-looking standard, new systems should be held to a higher standard
    - (but what is the upper bound, given that humans always make some mistakes?)

Additional Research

- Reproducibility: How much flexibility can be allowed in the test protocol?
  - Will variability in test participants experience due to labs in different geographic regions affect results?
  - Should we factor in older population or less educated population?
  - Benchmark thresholds are always tied to the demographics of the test participants to some extent
- Accessible voting system performance?