

**Usability Performance Benchmarks
For the Voluntary Voting System Guidelines**

Prepared at the direction of the HFP Subcommittee of the TGDC

August 17, 2007

This paper has been prepared by the National Institute of Standards and Technology at the direction of the HFP subcommittee of the Technical Guidelines Development Committee (TGDC). It may represent preliminary research findings and does not necessarily represent any policy positions of NIST or the TGDC.

The Technical Guidelines Development Committee is an advisory group to the Election Assistance Commission (EAC), which produces Voluntary Voting System Guidelines (VVSG). Both the TGDC and EAC were established by the Help America Vote Act of 2002. NIST serves as a technical adviser to the TGDC. The Human Factors and Privacy (HFP) Subcommittee of the TGDC has been established by the TGDC.

Overview

An accurate voting process—the casting, recording, and counting of votes—is a basic need for a democracy. To cast votes, a voter interacts with a voting system to record choices on a ballot. However, this interaction is not always straightforward. The voting system technology, the layout of the ballot, the contests themselves, or the instructions can sometimes be quite confusing. The usability of a voting system refers to the ease with which voters can interact with a voting system to record their choices as they intended.

The Technical Guidelines Development Committee (TGDC) intends to include requirements for voting systems to meet performance benchmarks for usability in its recommendations for the next version of the Voluntary Voting System Guidelines (VVSG). The goal of the new requirements is to improve the usability of the next generation of voting systems. Voting systems will be tested to see if they meet the benchmarks by test laboratories designated by the Election Assistance Commission. If a voting system meets or exceeds the benchmarks, then it is considered to have good usability. When using systems with good usability, voters will be able cast their votes more accurately and with less confusion. With these new performance requirements in place, the next generation of voting systems should have significantly improved usability.

Purpose and Scope

This paper describes those requirements, the benchmarks, and the work done under the direction of the Human Factors and Privacy (HFP) Subcommittee of the TGDC to develop both the requirements and the way systems will be tested to determine if they meet them.

The main result of this research is the development of a standard test methodology to measure whether a voting system meets usability performance benchmarks and proposed values for these benchmarks.

Usability can be informally described as the ease with which a user can operate, prepare inputs for, and interpret outputs of a system. The standard definition [ISO9241] of usability is “the 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.” For voting systems, usability measures the capability of voting systems to enable voters to cast votes as they intended, with few errors, quickly, and without frustration.

The usability requirements in this paper focus on accuracy as the measure of effectiveness of a voting system and set three benchmarks which all systems must meet.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

The HFP decided to also measure the length of time it took for test participants to complete their ballots as well as how confident they were that they had been able to make their vote choices as they intended. These measures are useful information especially to election officials who will be purchasing voting systems and can be related to the usability of the systems. Confidence was not shown to distinguish between the usability of the different systems; most test participants had similar good satisfaction and confidence levels for all systems tested. Accordingly, the TGDC decided that both efficiency and confidence should be reported, but not be used as requirements. No benchmarks were developed for these measures.

Design Requirements vs. Performance Requirements Using Benchmarks

The VVSG contains two types of usability requirements for voting systems. There are *design requirements* that reflect the expert judgment of usability and election professionals. For example, there are requirements for a minimum font size of 3.0 mm, standard colors, and complete instructions.

There are also *performance requirements*. To determine if a voting system meets these requirements, test laboratories use test participants voting in a controlled setting similar to an election to measure usability. They do this by measuring the capability of the voting system to enable those test participants to accurately cast votes.

Usability system performance requirements based on usability tests have two significant advantages over design requirements. First and foremost, *performance requirements directly address the “bottom-line” usability properties* of the system, such as how accurately can voters cast ballots, whereas design requirements do so only indirectly. Second, performance requirements are technology-independent – they provide impartial metrics of usability that are applicable across various types of voting systems: Direct Recording Electronic (DRE) systems, Electronic Ballot Markers (EBM), Precinct Count Optical Scanners (PCOS), etc. Because they are technology-independent, the use of performance requirements should allow voting system manufacturers to develop innovative interfaces without being overly constrained by design requirements.

A performance requirement for voting system usability has two components:

1. A reliable *test* for consistently assessing the usability of a voting system, and
2. A *benchmark*, a score or value that the voting system must achieve.

To assess whether the voting system meets the benchmark, the test method is tightly controlled with the test conducted in the same manner in the same environment and with each test participant given the same instructions on how to vote. Based on the accuracy of how all the test participants voted in each contest on the ballot, the voting system receives a score. The benchmark is the minimum “score” on the test that satisfies the performance requirement. If the system meets or exceeds the benchmark, then it “passes” the test and conforms to the requirement. These tests and benchmarks will become part of the conformance testing that Voting System Test Laboratories (VSTLs) must perform to

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

determine whether a voting system meets all the requirements in the VVSG and is thus considered sufficiently usable.

How the benchmarks were developed

The first part of the research included the development of a valid test method, that is, a test that could detect the types of errors that have been seen in prior research and one that is able to detect differences between types of voting systems. The work to develop the test method is described in detail later in this paper.

To develop the performance tests, a test ballot with 20 fabricated “contests”, including both election races and referenda, was prepared. This ballot was designed to be sufficiently complex to expose usability errors that have been reported in other voting system usability research and in other types of kiosk-based systems such as ATMs. Vendors with several different types of voting equipment were recruited to implement the ballot in the most usable way possible with their systems. The four (4) systems used in this research included a selection of DREs, EBMs, and PCOS.

Test participants were recruited from a specific set of requirements for age, gender, education, etc. Approximately 450 different test participants ‘voted’ in nine (9) tests on four (4) different voting systems. The demographics used for this research included an educated (college courses or degree) and younger set of test participants (25-54 years of age). These requirements were selected in part because if the test could detect usability differences and all the expected errors with this population, it would detect differences and errors with older and less educated populations as well. In addition, the research could then conclusively attribute the errors made by the test participants to poor voting system usability rather than to difficulties voters might have due to limited education or disabilities that may affect seniors or other groups of voters. Future research will include test participants who are more representative of eligible voters in the US. This will assist in further refining these benchmarks and will assure that the entire population is represented in the benchmark settings.

The test participants were given a written list of 28 instructions telling them how to vote in 20 contests and were given as much time as they needed to complete their votes. They were found to make a range of errors, such as skipping a contest, choosing a candidate adjacent to the candidate they intended to choose, and voting where no vote was intended. The testing established that the test method did identify measurable usability differences between systems and that repeated testing produced consistent results for the same system. These results are important, because they establish that the chosen testing methodology is both valid and repeatable when using different sets of test participants, and therefore is suitable for use by a test laboratory to determine if a system passes a usability conformance test.

Finally, the votes by the test participants were recorded and the errors counted and statistically analyzed. This data was used to derive benchmarks for three basic aspects of voting system usability: (1) how accurately test participants voted, (2) if they were able to

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

complete the voting process and successfully cast their ballots, and (3) the degree of variability among participants.

The Performance Measures

The HFP Subcommittee decided to create three benchmarks that measure basic aspects of the accuracy of voting systems:

1. **Total Completion Score:** the percentage of test participants who were able to complete the process of voting and cast their ballots so that their ballot choices were recorded by the system.
2. **Voter Inclusion Index:** a measurement that combines accuracy with the variability in the level of accuracy among individual test participants, and
3. **Perfect Ballot Index:** a measurement for detecting a systemic design problem that causes the same type of error by many test participants, by comparing the number of participants who cast a ballot without any errors to those that had at least one (1) error.

A **Base Accuracy Score**, the mean percentage of all ballot choices that are correctly cast by the test participants, is used in the calculation of the Voter Inclusion Index. This score, while not a benchmark itself, is critical to the calculation of any accuracy-related benchmark. The Voter Inclusion Index is a measure to identify those systems that, while achieving a high Base Accuracy Score, might still be less usable for a significant portion of the test participants. This measure distinguishes between systems which are consistently usable for participants versus those that have some participants making large numbers of errors. It ensures that the system is usable for all of the test participants.

Another dimension of accuracy is the number of errors by participants which might be caused by a particular design problem, even when there is a high accuracy for the voting system overall. The Perfect Ballot Index compares the number of cast ballots that are 100% correct with those that contain one or more errors. This measure helps to identify 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.

In summary, voting systems must achieve high Total Completion Scores, must have all voters voting with similarly high degrees of accuracy, and must achieve a high accuracy score, while also not allowing errors that are made by a large number of participants. Taken together, these benchmarks help ensure that if voters make mistakes, the mistakes are not due to systemic problems with the voting system interface.

The Benchmark Values Systems Must Achieve

As stated previously, a system must achieve passing scores (or benchmarks) for the measures of voting accuracy: Total Completion Score, Voter Inclusion Index, and Perfect Ballot Index.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

The derivation of these benchmarks required detailed statistical analyses, which is described in the main body of this paper. Based on that work, the benchmarks were then proposed by the HFP subcommittee for use in the VVSG usability requirements, to improve the usability of the next generation of voting systems. The TGDC will subsequently determine the exact benchmark levels. The current benchmarks were chosen such that several of the systems used for this research, which are currently in use in actual elections, would have difficulty meeting them.

No system is likely to pass the benchmark tests unless its Base Accuracy Score is above 90%. The Subcommittee's proposed benchmarks can be summarized as follows:

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*

Final Conclusions of this Research:

This research has established a standard test methodology to determine whether a voting system meets usability performance benchmarks.

The performance requirements provide impartial metrics that can be applied across various voting system technologies (DRE, EBM, PCOS, etc.).

Using performance tests as a reliable usability measuring tool requires a tightly defined test protocol: ballot, tasks, participants, environment, and metrics.

The testing protocol has strict testing controls to isolate and measure the effect of the voting system on usability, and not the effect of other variables. The testing protocol helps assure consistency in results; a voting system measured repeatedly should get, statistically, the same scores each time.

The testing protocol also allows comparability among different voting systems.

The results of performance testing are reasonably related to real-world performance, although this conclusion can be supported only indirectly.

Based on test results, benchmarks can be set to discriminate between systems that perform well and those that perform poorly.

It is a reasonable expectation that the application of performance tests will be a powerful tool to promote the development of demonstrably more usable voting systems.

See Appendix C for Frequently Asked Questions about this research.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

1. Introduction

The Help America Vote Act (HAVA) of 2002 mandated that the Election Assistance Commission (EAC), in consultation with the Director of the National Institute of Standards and Technology, submit a report on human factors, usability, and accessibility to Congress. The resulting EAC report included two recommendations for the development of usability performance benchmarks for voting systems:

- Develop voting system standards for usability that are performance-based, high-level (i.e., relatively independent of the technology), and specific (i.e., precise).
- Develop a valid, reliable process for usability conformance testing of voting products against the standards described in the recommendation above with agreed upon pass/fail requirements.

The Human Factors and Privacy (HFP) Subcommittee of the Technical Guidelines Development Committee, formed under HAVA, subsequently requested that NIST develop usability performance requirements for inclusion in the Voluntary Voting Systems Guidelines in Resolution #5-05 Human Performance-Based Standards and Usability Testing:

“The TGDC has concluded that voting systems requirements should be based, wherever possible, on human performance benchmarks for efficiency, accuracy or effectiveness, and voter confidence or satisfaction. This conclusion is based, in part, on the analysis in the NIST Report, *Improving the Usability and Accessibility of Voting Systems and Products* (NIST Special Publication 500-256).

Performance requirements should be preferred over design requirements. They should focus on the performance of the interface or interaction, rather than on the implementation details. ...

Conformance tests for performance requirements should be based on human performance tests conducted with human voters as the test participants. The TGDC also recognizes that this is a new approach to the development of usability standards for voting systems and will require some research to develop the human performance benchmarks and the test protocols. Therefore, the TGDC directs NIST to:

1. Create a roadmap for developing performance-based standards, based on the preliminary work done for drafting the standards described in Resolution # 4-05,
2. Develop human performance metrics for efficiency, accuracy, and voter satisfaction,
3. Develop the performance benchmarks based on human performance data gathered from measuring current state-of-the-art technology,
4. Develop a conformance test protocol for usability measurement of the benchmarks,
5. Validate the test protocol, and
6. Document test protocol.”

This report summarizes the research conducted to develop the test, metrics, and benchmarks in response to this resolution and describes the resulting performance-based requirements proposed for inclusion in the VVSG by the HFP. Supporting materials such as the test data and test materials can be found at <http://vote.nist.gov/benchmarks.htm> under “Usability Performance Benchmarks Supporting Materials.”

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

This research included:

- Defining a user-based test for measuring effectiveness, efficiency, and user satisfaction of voting systems
- Defining the metrics by which voting systems are evaluated
- Validating the test methodology
- Determining that the test protocol is valid and repeatable
- Setting performance benchmarks based on data collected by running the test on various typical voting systems.

2. The Problem: how to formulate testable performance requirements for the usability of voting systems?

The goal of the Usability section of the Voluntary Voting System Guidelines (VVSG) is to improve voting systems by setting requirements that will increase usability of voting system – the ability of voters to cast their ballots easily and correctly. This paper describes the research leading to the development of a test method for measuring the usability of a voting system. Applying this test method to current voting systems will give results that can then be used to determine benchmarks to separate those systems with good usability from those with poorer usability. As part of the VVSG, the test method and benchmarks will become part of the much larger set of requirements that will be applied by test laboratories to inform the EAC as to which voting system designs to certify as passing the usability section of the VVSG.

Usability engineering is the research and design process that ensures a product has good usability. There are two primary approaches that usability engineers employ to determine the usability of a system. The first is an evaluation by a usability expert who is experienced in identifying potential usability problems. The second approach is to set up a reasonably realistic test with representative users of the system and observe or measure how they actually interact with a system. The usability testing process can be used to uncover various types of problems with the interface or can be used to draw conclusions about a system's usability performance.

The VVSG contains requirements based on both approaches. There are *design requirements* that reflect the expert judgment of usability and election professionals and can be evaluated by an expert in usability and voting. For example, there are requirements for a minimum font size of 3.0 mm, conventional use of color, and complete instructions. There are also *performance requirements* that require usability testing with voters.

This paper describes draft performance requirements developed by the Human Factors and Privacy Subcommittee and explains how the requirements and associated performance test method, metrics and benchmarks were developed.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

A performance requirement needs two components before it can be tested: a reliable test method for measuring the system against the requirement and a benchmark. In the case of the usability performance requirements, the test method must be tightly controlled with representative voters using the voting system in a realistic election environment. Once a benchmark has been established, the system “passes” the test if its performance with the same test method is equal to or better than the benchmark established. The test methods and benchmarks will eventually be part of the testing that the Voting System Test Laboratories (VSTLs) will conduct to determine that a voting system meets all the requirements in the VVSG.

Every voter is different. Every election and associated ballot is different. So, how can a test for a voting system be developed to address whether the system is easy for voters to use effectively? To accomplish this, it must be shown that the test is *valid* – that is actually measures what it is supposed to measure. How can this be done in a test laboratory? How can we be certain that if a given voting system is tested several times with several sets of participants the pass or fail outcome will be the same, so that vendors, election officials, and voters will trust the results? This is accomplished by ensuring the test is *repeatable*. Repeatability is achieved by holding all other factors (ballot choices, environment, and characteristics of participants) as constant as possible.

Because of the controls that are required in a test environment, such test results, though, ***will not predict the actual performance of the voting system*** when used in a real election. It will predict the relative degree of usability of a system as measured against the benchmark values. However since the assessment is shown to be a valid measure of usability, a voting system that passes the test will do better than a system that fails the test in an election, generally speaking. This ensures that States will at the very least know that they have a system, if certified to the VVSG, with reasonable usability. The environment at the polling place can enhance or detract from that usability by how it is used or the environment it operates in. For example, good instructions and helpful poll workers offering assistance will improve usability, while long lines at the polls causing stress or poorly worded ballots may decrease usability.

This paper reports on the development of a performance test for the usability of voting systems. The performance test has six parts:

1. a well-defined test protocol that describes the number and characteristics of the voters participating in the test and how to conduct test,
2. a test ballot that is relatively complex to ensure the entire voting system is evaluated and significant errors detected,
3. instructions to the voters on exactly how to vote so that errors can be accurately counted,
4. a description of the test environment
5. a method of analyzing and reporting the results, and
6. the performance benchmarks and their associated threshold values.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

3. Defining the test

We developed a test to be used in conformance testing of voting systems that included measures of effectiveness, efficiency, and user satisfaction of voting systems, the three metrics that are standard for measuring usability. The standard definition [ISO9241] of usability is “the 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.” For voting systems, usability measures the capability of the voting system to enable voters to cast votes as they intended, with few errors, quickly, and without frustration. Based on the results of the studies described here, our benchmarks are effectiveness measures.

We call the test method the **Voting Performance Protocol (VPP)**. The VPP follows usability testing best practices [Dumas99]. The VPP is a *repeatable, controlled usability test* which will be used for conformance testing. The goal is not primarily to provide a “realistic” measure of voting behavior, but rather to isolate and measure the effect of various voting systems on user performance – and this requires that all other potential variables be held as constant as possible. See Section 9 for an overview of the VPP.

There were two challenges in defining the VPP:

1. The conformance test itself had to be as controlled and objective as possible so that staff from an accredited Voting System Test Laboratory (VSTL) could conduct the test, analyze the data, and obtain results that are reliable (that is, repeatable under exactly the same conditions and reproducible by another lab).
2. We needed to determine the number of test participants sufficient for repeatability of the conformance test and to allow for statistical analysis.

The test protocol also includes a *reference system*. The observed performance on this system is used to ensure that the data obtained from a system being tested is valid data for assessing the system’s performance against the benchmark values. See “Set Up” and “Procedure Validation” in Section 9.

This protocol includes:

- A ballot of moderate complexity (as measured against a survey of real ballots from across the United States)
- A ballot that includes options for straight party voting and write-in votes
- Instructions that require filling out almost the entire ballot with a specific choices to be made
- A scripted situation in which the test participants attempt to follow those instructions
- A specification for number and characteristics and number of test participants (voters)

The purpose of the NIST VPP is to support valid and repeatable metrics that can be used to judge the *relative* performance of various systems. This is sufficient for determining which systems have good usability and which do not and allows us to develop pass/fail usability performance requirements for the VVSG.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

3.1 The ballot and instructions

After conducting research into typical ballots from around the country, we designed a logical definition for the test ballot and specific voting instructions for filling out the ballot. The test ballot includes the following:

- Twenty contests, referenda, and ballot initiatives
- Contests at various depths of the government hierarchy (federal, state, local)
- Contests that were both partisan and nonpartisan
- Contests that were single member and multimember
- Retention races
- Constitutional amendments
- Ballot measures

Realistic but not real names were chosen for all candidates, fictitious party names were created using names of colors, and fictitious constitutional amendments and ballot measures were created. This ballot definition is shown in the Appendix A.

Voting system vendors who agreed to lend their systems for use in these tests were given this ballot definition and were asked to implement the ballot on their system to maximize their system's usability. To implement a ballot on a voting system typically requires a large amount of expertise in using the system's ballot definition language and understanding the capabilities of the voting system. A test laboratory would not have this expertise. If the ballot was poorly implemented, it could cause a system to fail. Asking vendors to implement the ballot as best they could ensures that the system design rather than a poorly implemented ballot is being tested. In actual elections, experts in the ballot definition, either vendors or state officials, implement the ballots.

The test participants were told to make specific voting choices for the test ballot. All instructions about the test process were given to the participants in writing, with no additional individual assistance offered. The following statement was read by the test administrator in response to any question from a participant:

“I'm sorry but I'm not allowed to help you once you start. If you are having difficulties you can try to finish. If you are stuck and cannot continue, you can stop if you wish.”

The complete set of instructions is shown in Appendix B. Various tasks were included in the instructions for the participants to represent realistic voting events. These included:

- Voting for names that appeared at various locations within a list of names
- Voting only a partial slate in a multimember contest
- Skipping specific elements of the ballot
- Write-in votes

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

By instructing the participants how to vote, the difference between the “intended” votes of the test participants and the actual votes that they cast can be determined. Accuracy of the cast ballot is calculated by counting the number of correct votes, 28 being a perfect score. Note that both the test ballot and the tasks were constructed to be complex enough to expose the different types of errors that would occur if a voting system design had poor usability.

3.2 The test participants

For conformance testing, enough participants are needed to provide reliable results that could be statistically analyzed, but not so many participants that conducting the test would become financially infeasible for the test laboratories. The number of participants needed was primarily driven by our need to calculate reasonably narrow 95% confidence intervals for the various metrics. Best practice in many other applications using a formula similar to that of one of our benchmarks (the Voter Inclusion Index) is to require a minimum of 100 participants so that a normal distribution can be assumed in the statistical analysis for determining a 95% confidence interval. This led to the need for at least 100 test participants for testing a system for conformance to the benchmarks. These calculations are explained in more detail later in the paper.

One of the more complicated issues for development of the VPP concerns the diversity of the test participants to be recruited. Described below are several issues considered in developing the demographics of test participants. (Also, see Section 5.1 for a description of the test participant profile actually used for the determination of the benchmark values.)

Consistency across test participant groups: Each group of test participants recruited for a test must be as similar as possible to the groups used for other tests. To do otherwise would introduce variations in test results threatening the test’s repeatability. The approach used in this protocol was to specify approximate ranges for attribute values, e.g. “approximately 60% female.”

Attributes to control: The simple explanation is that one must control the attributes that affect performance, but not those that do not. The variables chosen to control are those that are considered in typical best practice usability testing of software interfaces. The current version of the profile calls for controlling gender, race, education, and age. Other variables have not been considered because they haven’t been found to affect performance in many other software user interfaces.

Controlling for region presents certain unique difficulties. It is possible, but unlikely, that region could affect performance because of voter experience. Voters from certain states might have more or less experience with various kinds of voting tasks, e.g., straight party voting or complex ballots, and equipment, e.g., manually-marked paper ballots (MMPB), lever machines, or Direct Recording Electronic (DRE) systems. On the other hand, as a practical matter, it would be difficult for test laboratories to recruit a broad cross-section of voters from various regions for each test. The tests performed thus far were with test

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

participants from a region that has a very transient population (Virginia, Maryland, and the District of Columbia) so there was some variation in experience and this did not seem to affect the repeatability. If region does *not* have a significant effect on performance, then the problem goes away. The actual effect of region on performance is being addressed with further research.

A final point on attributes: the more finely we specify attributes for test participants, the more difficult it becomes to recruit a consistent group. As a practical matter, the size of the recruited group has to become considerably larger in order to “cover” the various characteristics. In addition, the cost of recruiting a group with a number of very specific characteristics can be prohibitively high. The goal of this test was to define a reasonable set of attributes to ensure repeatability but not so specific to be cost prohibitive. With additional research, we will continue to determine where more flexibility in the characteristics is possible.

Representativeness: Should the values for the controlled attributes of the test participants closely reflect the profile of the actual voting population? While it might seem logical that a sample of test participants that precisely matches the U.S. population of voters or eligible voters would make the ideal sample population, it is not necessary for our goal and would be cost prohibitive. It also does not contribute to testing that would model voters in real elections. Elections are held at the state and local level, with ballots that are designed according to state election laws. The voters participating in a specific election do not match the demographic of the U.S. population. The purpose of the demographic profile in the VPP is to provide a consistent set of test participants whose experiences will discriminate relative differences in usability among various voting systems at an economically feasible cost. Further, the test ballot is designed to represent many types of voting behavior some of which might not occur in a local jurisdiction because of local election law.

A conformance test only requires that the performance of a system be compared against the benchmark values developed using similar populations. The participants are not voting for their own choices, they are participating in a controlled laboratory test following written directions and voting for a prescribed set of candidates and referendum choices and doing so without any assistance.

Revealing Errors: If the test was too simple and all participants were able to use the system without errors, there would be no way to determine a difference in usability between the systems. Clearly, we did want to recruit a group of test participants who are potential voters and who would make a variety of errors in systems. We had initially assumed that potentially “vulnerable” categories of voters, such as the elderly, or those with little education, were needed to uncover usability problems. This was not the case: even a population that could be considered “above average” in some dimensions made enough errors and a large variety of errors in the tests so that differences in usability performance between systems were easily detected. Future work will ensure that the usability needs of seniors, the disabled, and other “at risk” groups are represented in the results reported to the EAC.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Benchmarks: The pass/fail benchmarks values are a function of the test participants' attributes. For example, a group of participants with 25% over the age of 70 is likely to generate different results from a group with only 10%. The benchmarks discussed below in section 5 are dependent on the characteristics of the participants used in the benchmark tests. Changing those characteristics in the VPP would possibly result in different benchmark values for the same test, but the test would still be valid.

Voters with disabilities: The current protocol addresses performance testing only for "conventional" systems; voters with disabilities have not yet been included the VPP profile. Future research will include testing accessible voting systems with specific populations of people with disabilities.

3.3 The test environment and voting equipment

As these are controlled studies representing conformance tests, they were performed in a usability test lab so that the environmental conditions such as lighting could be held constant. Future versions of the VPP will specify the testing environment (and other steps) as comprehensively as possible so that the test laboratories can conduct testing under nearly identical conditions.

The four (4) voting systems used in these tests were systems certified under the Voting System Standards 2002 (VSS 02) and used in at least one election, but, of course, the test method is designed to be applicable to any voting system. They included a selection of Direct Recording Electronic (DRE) systems, Electronic Ballot Markers (EBM), and Precinct Count Optical Scanners (PCOS). Vendors were asked to implement the VPP test ballot for their systems to best demonstrate the usability of their equipment, as described in Section 3.1.

4. Defining how the systems were measured

Performance metrics for effectiveness, efficiency, and user satisfaction were defined as follows:

1. Effectiveness was measured as:
 - a. **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. Failure to cast a ballot might involve problems such as a voter simply "giving up" during the voting session because of an inability to operate the system, or a mistaken belief that the casting has been successful executed.
 - b. **Voter Inclusion Index:** A measure of overall voting accuracy that uses a Base Accuracy Score and the standard deviation. The **Base Accuracy Score** is the mean of the percentage of all ballot choices that are correctly

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

cast by each of the test participants. Each voter is given 28 “voting opportunities” within the ballot. The number of these that are correctly performed divided by 28 yields that voter’s base accuracy score. The Base Accuracy Score, while useful for measuring the “accuracy” of voting, does not indicate whether all participants, are able to vote with approximately equal accuracy. The Base Accuracy Score is used to compute the Voter Inclusion Index which *does* distinguish between systems which are consistently usable for participants versus those that have some participants making large numbers of errors. We calculate the Voter Inclusion Index (VII) as follows:

$$VII = \frac{\bar{X} - LSL}{3S}$$

where LSL is a lower specification limit, \bar{X} is the mean of the base accuracy scores of each of the participants, and S is the standard deviation. We used 85% as our LSL. The LSL acts as an absolute lower limit for the accuracy rate. Any system with an accuracy rate at 85% or below gets a 0 or negative VII and fails. The choice of 85% is somewhat arbitrary: it serves to spread out the range of the VII values for different systems so that there is a larger separation on the measurement scale between systems. The higher the value of the index, the better the performance of the system. The VII is a type of “process capability index”. A complete description of process capability and how to use and calculate capability indices can be found in

<http://www.itl.nist.gov/div898/handbook/pmc/section1/pmc16.htm>

To illustrate how the Voter Inclusion Index is a measure of the amount of variability between test participants in how accurately they vote and how it helps to identify those systems that, while achieving a high Base Accuracy Score, still produce errors within a significant portion of the test participants, we give an example. Suppose that two different voting systems have the same accuracy score of 96%. However, one system achieves this score for almost every test participant, while the second system achieves this score with 80 participants achieving 100% accuracy and another 20 achieving 80% accuracy. The second system would have a lower Voter Inclusion Index even though its Base Accuracy Score was identical to the first system.

c. Perfect Ballot Index: 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

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

problem, causing a similar error by the participants. The higher the value of the index, the better the performance of the system.

2. Efficiency was measured as **Average Voting Session Time** – mean time taken per test participants to complete the process of activating, filling out, and casting the ballot.
3. User Satisfaction was first measured using a Likert scale satisfaction questionnaire. In a later test, we measured user satisfaction as **Average Voter Confidence** – the mean confidence level expressed by the voters that they believed they voted correctly and the system successfully recorded their votes. This was based on a confidence question we developed specifically for this test.

Additional general information about measuring usability and related statistical methods is available at: <http://www.measuringusability.com>.

5. Setting the benchmarks

Four different voting systems were tested to determine possible values for benchmarks for each of the three effectiveness measures. Four sets of approximately 50 individuals were recruited, one for each of the four systems.

This research was conducted between May and July of 2007, using four voting systems that had been certified to the 2002 VSS and used in at least one actual election. The systems were from four different vendors and are representative of current systems and technology. As mentioned earlier, these included a selection of DRE, EBM, and PCOS systems. Test labs had not yet begun testing to the VVSG 2005 and so there were no available systems known to conform to VVSG 2005.

5.1 Test participant demographics

For these initial benchmark tests, the demographics chosen were based on the ability to recruit a consistent set of participants for multiple tests.

Our validity and repeatability studies performed prior to the benchmark performance research, and explained later in this paper, demonstrated that even this relatively narrow set of demographic characteristics did produce sufficient data for our analysis. Our test participants closely resembled the following demographic criteria:

- 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

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

While the geographic distribution also seems rather limited, it should be noted that this is a very transient population with a variety of prior voting experiences. As noted earlier, future research will be performed to compare how this demographic relates to other voter populations. Technical discussion on this point is above in Section 3.2. The details of the exact demographics can be found in <http://vote.nist.gov/benchmarks.htm> under “Usability Performance Benchmarks Supporting Materials.”

5.2 Performance data used to set the benchmarks

Data collected from this test produced the results in Tables 1 and 2 for Total Completion Rate, Base Accuracy Score, Voter Inclusion Index, and Perfect Ballot Index.

Metrics can only be calculated based on the data for the test participants; they do not measure the “true value” directly and the results are only accurate to a certain level of confidence. To incorporate this level of confidence, a *confidence interval* is calculated around the data collected. In our case, we are using a 95% confidence interval. This confidence interval tells us that, by definition, we are 95% confident that the “true” value of the system for a given metric lies within the confidence interval (CI). The tables below show results at the 95% Confidence Interval (CI) determined for the completion score and the two indices. To determine if a system passes the test, we compare the system’s 95% confidence interval against the benchmark value. If the entire CI falls below the benchmark, we are 95% confident that the system does not pass the benchmark. Since we are only 95% confident, this also means that in a very small number of cases, the test might falsely fail a system. (For a VSTL, this means that if a voting system fails by a very small margin, the vendor might ask that the test be performed again.)

Table 1 Summary Benchmark Performance Data by Voting System: Total Completion Score, Base Accuracy Score, and Voter Inclusion Index

	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

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

**Table 2 Summary Benchmark Performance Data by Voting System:
Perfect Ballot Index, Time, and Confidence**

	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 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

Based on the data collected in these tests, it is possible to derive benchmarks for requirements that must be met in conformance tests for the VVSG. The suggested benchmarks were chosen based on the assumption that they should not be so stringent as to fail all systems, but not so lenient as to pass all systems. Systems with significantly lower performance as measured by the benchmarks should fail. The TGDC will use this data to help make a decision on the values of the benchmarks. The proposed values here are just suggestions based on current systems. The TGDC may decide, as a policy, to choose benchmarks that will push system design towards higher levels of usability.

5.3 Total Completion Score Benchmark

The calculation of the confidence intervals for the Total Completion Score Benchmark is based on the Adjusted Wald Method. The Total Completion Scores for all the systems tested were relatively high and showed that it is possible for a system to achieve a Total Completion Score of nearly 100%. It seems reasonable to set the value for this rate at a fairly high level for the laboratory conformance test since this value represents all but the few individuals who could not finish once they started the voting task.

The ranges of values obtained in this set of tests suggest setting the benchmark for the Total Completion Score at 98%.

As an example of what this means and how the confidence interval is used, suppose there were 100 participants who attempted the voting test, that system would pass the

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

benchmark if at least 95 participants completed the test, because when 95 participants succeed the resulting 95% confidence interval is [.8854, .9813] which includes the 98% benchmark. However, if more than five participants fail to complete the test, then the system would fail on this measure since the entire 95% confidence interval is below the 98% benchmark.¹

5.4 Voter Inclusion Index

The Voter Inclusion Index levels in this test showed a wide range of results, with two of the systems significantly below the other two systems. These differences in the index were not the result of significant differences in the Base Accuracy Scores (which were in the 90% range for all four systems), but in the large differences in the standard deviations. A high standard deviation reflects high variability among the performance of individual test participants. A lower standard deviation implies a more equal level of performance among participants. The mean and standard deviation of the Base Accuracy Score for each of the machines tested are shown in Table 1 above.

A conformance level for Voter Inclusion Index with a goal of a mean of at least 92.5% and a standard deviation of no more than 7% was calculated. Using a lower specification limit of 85%, the calculated cut off point for the Voter Inclusion Index is .35, and this is how the suggested benchmark was chosen.

The data suggests setting the benchmark for the Voter Inclusion Index at .35.

With an accuracy index benchmark of .35, two of the systems pass and two fail.

Note that the purpose of this testing was not to diagnose problems. The test administrators did observe that some test participant confusion and errors indicated that improvements (and better results) could be obtained in all of these designs.

5.5 Perfect Ballot Index

The Perfect Ballot Index is calculated, for example for System A, as

$r = \frac{29}{50 - 29} = \frac{29}{21} = \frac{p}{1 - p} = 1.38$, where $p = \frac{29}{50}$. However, the values of the rate in the tables were calculated based on the Adjusted Wald method. Here for System A based on that, $r = 1.35$ instead of 1.38. The transformation is

$r = \frac{p}{1 - p} \cong f(p)$. Since the derivative $\frac{dr}{dp} = \frac{1}{(1 - p)^2} > 0$, the random variable r is an increasing function of the random variable p . Thus,

¹ The reader can perform the calculations for this example at <http://www.measuringusability.com/wald.htm> using the calculator and the Adjusted Wald confidence interval.

$$P(a < p < b) = P(f(a) < r < f(b)) = 0.95$$

The values of $f(a)$ and $f(b)$ are the limits of the 95 % confidence interval of r . For System A in Table 2, $a = 0.442$ and $b = 0.706$. The corresponding $f(a) = 0.79$ and $f(b) = 2.40$.

The participants' voting ballots with no errors was lower than might be expected in all systems tested. There are at least two possible reasons for this finding: first, the test itself represents a moderately difficult task in which participants are required to match values from the instructions for 20 separate contests. Second, this rate represents any error out of the 28 specific votes (or non-votes) expected. Many participants had only a single error on their ballots.

The ranges of values obtained in this set of tests suggest setting the benchmark for the Perfect Ballot Index at 2.33.

The calculation of the confidence intervals is based on the Adjusted Wald Method similar to the calculation for the Total Completion Score confidence intervals.

At this value, the three of the four systems tested would pass. Note that retesting these systems with 100 (rather than 40-50) participants would narrow the confidence intervals and therefore make the test more accurate, that is, it would generate a narrower confidence interval. Note that for the statistical reasons outlined earlier, the VPP will require at least 100 test participants for the conformance tests performed by the test laboratories.

5.6 Why these benchmark values?

To reiterate, it is important to note that the HFP subcommittee chose these benchmarks so that they are achievable by some systems, but not so low as to be trivially easy to meet with any of the current voting system implementations. The TGDC may want to consider setting them somewhat higher to encourage further usability improvements for systems that will be certified to this new version of the VVSG. The purpose of these technical findings is to provide some useful guidance to the TGDC. The degree of stringency to be reflected in the benchmarks is basically a policy question.

5.7 Voting time, satisfaction, and confidence

In analyzing the research data on time, two points emerged: 1) the systems tested showed a wide variability in the time required to complete the same ballot (note the wide range of average benchmark performance as well as wide ranges for individuals' performances on each system), and 2) longer time was not always correlated effectiveness. While it would appear to be technically feasible to construct a benchmark based on time alone, the policy question is whether to use time as a pass/fail criterion, given that poor time performance affects throughput but does not necessarily correlate with voter effectiveness.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

We attempted to collect confidence data in two ways. In earlier tests satisfaction was tested, using a questionnaire with a 5-point scale and six (6) questions. In later tests we asked a single question about the participants’ confidence that they were able to use the system correctly. Neither measurement appeared to be sensitive enough to be used for a performance benchmark. Most test participants had a high degree of satisfaction and confidence regardless of their effectiveness.

Because of these results, we modified the satisfaction question for the benchmark tests, based on cognitive interviews with test participants. This is a technique used to ensure that test participants understand the question being asked. Our goal was to develop a single question that captured two elements: the perception of the participants’ ability to use the system and the confidence of the participants in the system itself within the larger context of the voting process.

This is the revised confidence question that we used during Benchmark testing.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
	1	2	3	4	5
If I had to use this machine in a real election, I’d be confident in my ability to enter my votes and I’d trust this machine.					

Note that this question also generated consistent results regardless of the system used even though performance differences were measured.

The TGDC has decided not to make time or confidence pass/fail requirements in the VVSG. We suggest that Average Voting Session Time and Average Voter Confidence be tested and reported by the Voting System Test Laboratories (VSTLs) because election officials may find them helpful for determining the number of vote stations needed and general attitudes held by voters towards the systems.

5.8 Sufficient number of test participants

Based on the above studies, 100 test participants per test is adequate. Not only does this number make it statistically possible to do the Voter Inclusion Index calculation, based on the studies described in this paper with sets of 50 test participants, more than 50 test participants appears to be sufficient. With 100 participants, the 95% confidence intervals are more narrow which means we are getting results very close to the true performance of a voting system with a high degree of confidence. The Voter Inclusion Index is a type of process capability index and the best practice in other domains suggests that at 100 data points, a normal distribution can be assumed, thus simplifying the confidence interval calculations. The cost to recruit 100 participants and run the test is reasonable and within the range of the typical cost for conformance testing by a VSTL.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

6. Validating the test methodology

As a first step prior to the benchmark testing, the test methodology was validated to determine whether it could detect performance differences between voting systems and whether the differences measured met expectations.

6.1 Running the test for validation

We had two systems (we will call them A and B), which prior research [MIT01] had shown would be expected to have different performance on at least some of our measures. Twenty four (24) participants were asked to use System A and twenty- three (23) participants were asked to use System B using the test ballot and procedures we had defined. This was a sufficient number of test participants to show validity—we did see statistically significant differences between systems where we expected to see them.

Data collected from this test produced the results shown in Table 3 for the system’s Voter Inclusion Index, Perfect Ballot Index, voting time, and satisfaction using the questionnaire we developed. (We did not collect data for Total Completion Score as it was not necessary for demonstrating the validity of the test.) Data from participants who did not complete the voting task are not represented in any of the tables for the Voter Inclusion Index, Perfect Ballot Index, timing data, or satisfaction. That data is represented only in the Total Completion Score.

Table 3: Summary Performance Data by System for Validity Test

	Total Completion Score	Base Accuracy Score (%)	Voter Inclusion Index	Perfect Ballot Index	Voting Time (secs)	Satisfaction (%)
	95% Confidence Intervals	Mean, Standard Deviation	95% Confidence Intervals	95% Confidence Intervals	Mean, Standard Deviation	Mean, Standard Deviation
System A	N/A	42.6-78.9	.85-1.62	0.74 – 3.74	402.56, 103.08	80.7, 15.2
System B	N/A	22.1-59.3	0-.24	0.28 – 1.46	431.65, 114.30	73.6, 16.3

6.2 How we analyzed the test results

We used statistical analyses (a binomial calculation called Adjusted Wald) for Perfect Ballot Index confidence intervals. For the Voter Inclusion Index, with a large enough set of test participants, we can assume an asymptotic normal distribution to calculate the 95% confidence intervals. By definition, we have 95% confidence that the true value (well-defined, but unknown) for the system is contained in the interval. These calculations showed us that we *are* able to detect differences in effectiveness between the two systems.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

For Perfect Ballot Index, of the 24 participants who attempted to vote using System A, 15 were able to cast the ballot without any deviation from the results expected (62.5%) and, of the 23 participants using System B, 9 were able to cast the ballot correctly (39%). The 95% confidence interval for System A's Base Accuracy Score is between 42.6% and 78.9% and for System B between 22.1% and 59.3%. Though there is overlap between these two ranges, this is likely due to the small number of participants in this evaluation. The data were in the direction expected based on expert review and prior research, such as [Greene06] and [Conrad06] with System A showing higher effectiveness.

For the Voter Inclusion Index, assuming a Lower Specification Limit of 85%, System A was measured at 1.24 and System B at 0.10. Calculating a 95% confidence interval for the Voter Inclusion Index gives us a range of .85 to 1.62 for the System A, and 0 to .24 for System B. Thus, the index also showed a statistically significant difference in the effectiveness between the two machines.

As an additional check, a Mann-Whitney test was used to check for statistically significant differences between the simple accuracy rates (as opposed to index) of the two systems.² Using this calculation, a statistically significant difference (p-value<0.05) was seen, with System A performing better than System B.

A statistical t-test was used to compare time data between the two systems. This analysis of the data showed no statistically significant difference between the two systems though System A had a slighter faster mean time than the System B.

A t-test was also used to compare the satisfaction data collected from the satisfaction questionnaire. Again, there was no statistically significant difference in the data though the satisfaction data showed somewhat better mean scores for System A than for System B.

The results of this testing indicated that the test was valid, in the sense that we were able to "tell the difference" between the two systems and it was what we expected. In this initial test, it was only effectiveness that could be used to discriminate between these two. However, as we saw earlier, the benchmark tests, performed on a wider variety of systems, showed significant differences in efficiency. The satisfaction questionnaire did not show statistically significant differences in the validation or in any of the repeatability tests, so we are not using it as a benchmark and we simplified it to a single question about confidence.

So far, none of our tests have shown strong correlations among the three types of metrics, e.g., effectiveness and efficiency do not appear to be strongly linked. Faster time does not necessarily imply accuracy or vice versa.

² The Mann-Whitney test, a nonparametric statistical analysis approach, was used because the performance data was not normally distributed.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

7. Determining whether the test protocol result is repeatable

Additional tests were conducted to determine if the test was repeatable – that the test returned results with no statistically significant difference each time it was conducted by the same test administrators on the *same system*.

Four tests of the same system were conducted. Only the set of participants changed. Between 44 and 50 individuals participated in each of these four tests. Data collected from this test produced results shown in Table 4.

Table 4 Summary Repeatability Performance Data by Test

Repeat-ability Test	Total Completion Score (%)	Base Accuracy Score (%)	Voter Inclusion Index	Perfect Ballot Index	Voting Time (secs)
	95% Confidence Intervals	Mean, Standard Deviation	Confidence Intervals	95% Confidence Intervals	Mean, Standard Deviation
Test 1	87.4-100	95.2, 14.5	.11-.33	1.15 – 4.03	695.4, 224.0
Test 2	88.3-100	92.5, 17.3	.05-.24	0.68 – 2.07	662.0, 245.0
Test 3	77.8-96.0	92.3, 16.3	.05-.25	0.46 – 1.49	691.0, 206.8
Test 4	86.3-99.7	95.2, 11.5	.19-.41	0.79 – 2.40	633.8, 166.5

- Total Completion Scores: For these four tests, 44 of 45, 48 of 49, 44 of 49, and 50 of 52 participants finished voting. The confidence intervals overlap indicating that the data are repeatable.
- The Voter Inclusion Index confidence intervals and Base Accuracy Scores for these four tests show sufficient overlap to indicate that these are consistent results
- Perfect Ballot Index: Participants who were able to complete voting were the smaller number in the Total Completion Score data (first bullet). For these four tests, the number of correct ballots as compared to cast ballots was 30 of 44, 26 of 48, 20 of 44, and 29 of 50. Again, the confidence interval ranges overlap indicating that the data are repeatable.
- The voting time was also analyzed and shown to be consistent for all tests (p-value >0.05).

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

The Mann-Whitney test was again used to determine if the accuracy rates of the four tests were consistent and they were (p-value > 0.05) except for Test 1 compared to the Test 3 – and these two tests were only marginally different.

8. The performance requirements in the VVSG

The HFP Subcommittee has proposed the following requirements for the VVSG. They are based on the benchmark testing described above in Section 5. These are subject to change as the TGDC determines stringency and as the additional research is conducted. They are:

Total Completion Score Performance

The system shall achieve a Total Completion Score of at least 98% as measured by the VPP.

Perfect Ballot Index Performance

The system shall achieve a Perfect Ballot Index of at least 2.33 as measured by the VPP.

Voter Inclusion Index Performance

The system shall achieve a Voter Inclusion Index of at least 0.35 as measured by the VPP.

Reporting requirements in VVSG, Volume IV include reporting the above results and:

Voting Session Time

The test laboratory shall report the average voting session time, as measured by the NIST VPP.

Average Voter Confidence

The test lab shall report the average voter confidence, as measured by the NIST VPP.

9. Voter Performance Protocol (VPP)

This is an overview of the VPP used for these tests. Eventually intended for the VSTLs, this user test protocol will be described in much greater detail as the test methods in support of the VVSG are developed.

- 1. Ballot onto voting system.** The voting system vendor is responsible for putting the test ballot specification onto the system to be tested.
- 2. Set up.** Two systems are set up for each test: the system being tested and *reference system* whose level of performance has been previously established. As we will see below, this is a way to calibrate the lab testing process and ensure that the test procedures have been followed properly

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

- 3. Participants and their demographics.** Participants are recruited for the test based on specific demographic characteristics of age, education, race, gender, and location. At least 100 are needed for a test.
- 4. Test environment.** The voting systems are installed in a lab under realistic but controlled conditions that make it easy for participant use (e.g., presence of a chair, appropriate lighting).
- 5. Method.**
 - a.** Participants arrive at the test lab, fill out the consent form, and wait in a room separate from the voting systems until it is their turn to participate. They are provided with the overview of the test procedures. They are told they will be asked to enter specific votes provided to them, and that the test facilitator will not be able to assist them once they have started.
 - b.** When it is their turn to participate, a test facilitator shows them the voting system to which they are assigned (the one being tested or the reference system) and gives them the full voting instructions. The participants are also given any additional materials needed to use the specific voting system (e.g., access code, paper ballot, voting card) but they are not told how to use them.
 - c.** To minimize interference in the measurement of usability, once the participant has begun the test, the facilitator's interaction with them is limited to the following statement: "I'm sorry but I'm not allowed to help you once you start. If you are having difficulties you can try to finish. If you are stuck and cannot continue, you can stop if you wish."
 - d.** Behind a mirrored glass, observers time the participants as they complete the voting task.
 - e.** After the participant completes casting the ballot or stops, the facilitator directs the participant to a third room where another individual administers the post-test questions and compensates them for their participation.
- 6. Data gathering.** A tally is kept of all participants who fail to complete the voting task either by leaving prior to completion or by leaving, believing they had completed the task but had not. For all participants that complete the voting task, the ballot data as counted by the system is entered into an Excel spreadsheet where they are tallied to determine the numbers of correct votes and the number of ballots cast without any errors.
- 7. Procedure Validation.** From these data, a comparison is made between the values obtained from the reference system and from previous values obtained for the same system. If the comparison shows the data to be consistent, the test is considered valid and the analysis is performed on the system being tested.
- 8. Data Analysis.** The data from the system being tested is converted into 95% confidence intervals for the three measures of interest (Total Completion Score, Voter Inclusion Index, and Perfect Ballot Index). Additional data, such as timing,

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

post-test questions, and demographic information, is entered into separate Excel spreadsheets for reporting.

10. Next steps

The research described in this white paper shows the validity and repeatability of the proposed testing methodology and provides data for determining benchmarks that specify usability performance requirements for voting systems. This methodology is technology-independent. Any type of voting system that can implement the test ballot can be evaluated for conformance to these requirements.

To ensure that this methodology can be reproduced in any VSTL, we will be performing research to determine how precisely the test conditions must be followed as part of our work on VVSG test method development. An open issue is that of demographics, the degree to which test participant characteristics must be controlled. For example, are there regional differences that would produce different results (e.g. do people in Chicago or Denver perform differently?). This will require conducting additional benchmark tests in different regions of the country. Future research will also include targeted testing of seniors and less educated voters, as well as testing of accessible voting systems with specific populations of people with disabilities to see if the current benchmarks will continue to apply.

References

[Greene06] Greene, K. K., Byrne, M. D., & Everett, S. P. A comparison of usability between voting methods. Proceedings of the 2006 USENIX/ACCURATE Electronic Voting Technology Workshop. Vancouver, BC, Canada.

[Conrad06] Conrad, F. G., Lewis, B., Peytcheva, E., Traugott, M., Hanmer, M. J., Herrnson, P. S., et al. (2006). The usability of electronic voting systems: Results from a laboratory study. Paper presented at the Midwest Political Science Association, Chicago, IL. April 2006.

[Dumas99] Dumas, J. and Redish, J.C. (1999). *A Practical Guide to Usability Testing*. Portland, OR: Intellect.

[Herrnson] “The Promise and Pitfalls of Electronic Voting Results from a Usability Field Test” by Herrnson, Niemi et al.

[ISO9241] ISO 9241, Part 11, “Ergonomic requirements for office work with visual display terminals”

[MIT01] “Residual Votes Attributable to Technology” at http://vote.caltech.edu/media/documents/wps/vtp_wp2.pdf

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

[Traugott] “The Impact of voting Systems on Residual Votes, Incomplete Ballots, and Other Measures of Voting Behavior” by Traugott et al

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Appendix A

Test Ballot Definition

Information applicable to whole ballot

Date and Time	2004-nov-02, 7:00 AM to 8:00 PM
State	Maryland
County	Madison
Party Line Voting Method	Enabled for partisan contests

Information applicable to every contest

Full-term or partial-term election	Full-term
Voting Method	Simple vote for N candidate(s) - (i.e. no ranked voting)

- **Contest #0:**

Title of Contest	Straight Party Vote
Partisanship	Partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	1
Maximum Write-in Votes Allowed	0

- **Option #0.1:** Blue
- **Option #0.2:** Yellow
- **Option #0.3:** Purple
- **Option #0.4:** Orange
- **Option #0.5:** Pink
- **Option #0.6:** Gold
- **Option #0.7:** Gray
- **Option #0.8:** Aqua
- **Option #0.9:** Brown

- **Contest #1:**

Title of Office	President and Vice-President of the United States
-----------------	---------------------------------------------------

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

District of Office	United States
Partisanship	Partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	1
Maximum Write-in Votes Allowed	0

- **Candidate #1.1:** Joseph Barchi and Joseph Hallaren / Blue
- **Candidate #1.2:** Adam Cramer and Greg Vuocolo / Yellow
- **Candidate #1.3:** Daniel Court and Amy Blumhardt / Purple
- **Candidate #1.4:** Alvin Boone and James Lian / Orange
- **Candidate #1.5:** Austin Hildebrand-MacDougall and James Garritty / Pink
- **Candidate #1.6:** Martin Patterson and Clay Lariviere / Gold
- **Candidate #1.7:** Elizabeth Harp and Antoine Jefferson / Gray
- **Candidate #1.8:** Charles Layne and Andrew Kowalski / Aqua
- **Candidate #1.9:** Marzena Pazgier and Welton Phelps / Brown

- **Contest #2:**

Title of Office	US Senate
District of Office	Statewide
Partisanship	Partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	1
Maximum Write-in Votes Allowed	1

- **Candidate #2.1:** Dennis Weiford / Blue
- **Candidate #2.2:** Lloyd Garriss / Yellow
- **Candidate #2.3:** Sylvia Wentworth-Farthington / Purple
- **Candidate #2.4:** John Hewetson / Orange
- **Candidate #2.5:** Victor Martinez / Pink
- **Candidate #2.6:** Heather Portier / Gold
- **Candidate #2.7:** David Platt / Gray

- **Contest #3:**

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

Title of Office	US Representative
District of Office	6th Congressional District
Partisanship	Partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	1
Maximum Write-in Votes Allowed	1

- **Candidate #3.1:** Brad Plunkard / Blue
- **Candidate #3.2:** Bruce Reeder / Yellow
- **Candidate #3.3:** Brad Schott / Purple
- **Candidate #3.4:** Glen Tawney / Orange
- **Candidate #3.5:** Carroll Forrest / Pink

- **Contest #4:**

Title of Office	Governor
District of Office	Statewide
Partisanship	Partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	1
Maximum Write-in Votes Allowed	1

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

Candidate #4.1: Charlene Franz / Blue

- **Candidate #4.2:** Gerard Harris / Yellow
- **Candidate #4.3:** Linda Bargmann / Purple
- **Candidate #4.4:** Barbara Adcock / Orange
- **Candidate #4.5:** Carrie Steel-Loy / Pink
- **Candidate #4.6:** Frederick Sharp / Gold
- **Candidate #4.7:** Alex Wallace / Gray
- **Candidate #4.8:** Barbara Williams / Aqua
- **Candidate #4.9:** Althea Sharp / Brown
- **Candidate #4.10:** Douglas Alpern / Independent
- **Candidate #4.11:** Ann Windbeck / Independent
- **Candidate #4.12:** Mike Greher / Independent
- **Candidate #4.13:** Patricia Alexander / Independent
- **Candidate #4.14:** Kenneth Mitchell / Independent
- **Candidate #4.15:** Stan Lee / Independent
- **Candidate #4.16:** Henry Ash / Independent
- **Candidate #4.17:** Karen Kennedy / Independent
- **Candidate #4.18:** Van Jackson / Independent
- **Candidate #4.19:** Debbie Brown / Independent
- **Candidate #4.20:** Joseph Teller / Independent
- **Candidate #4.21:** Greg Ward / Independent
- **Candidate #4.22:** Lou Murphy / Independent
- **Candidate #4.23:** Jane Newman / Independent
- **Candidate #4.24:** Jack Callanann / Independent
- **Candidate #4.25:** Esther York / Independent
- **Candidate #4.26:** Glen Chandler / Independent
- **Candidate #4.27:** Marcia Colgate / Independent
- **Candidate #4.28:** Leslie Porter / Independent
- **Candidate #4.29:** Molly Dalton / Independent
- **Candidate #4.30:** David Davis / Independent
- **Candidate #4.31:** May Peterson / Independent
- **Candidate #4.32:** Patricia Dawkins / Independent
- **Candidate #4.33:** Suzanne Adams / Independent
- **Candidate #4.34:** Mary Miller / Independent
- **Candidate #4.35:** Rosalind Leigh / Independent
- **Candidate #4.36:** Elaine Henry / Independent
- **Candidate #4.37:** Gail Moses / Independent
- **Candidate #4.38:** Daniel Jones / Independent
- **Candidate #4.39:** Don Maybee / Independent
- **Candidate #4.40:** Lillian Cohen / Independent
- **Candidate #4.41:** Richard Mitchell / Independent
- **Candidate #4.42:** Pat York / Independent
- **Candidate #4.43:** Linda Rappaport / Independent
- **Candidate #4.44:** Mike Porter / Independent
- **Candidate #4.45:** Margaret Sharp / Independent
- **Candidate #4.46:** Cathy Steele / Independent
- **Candidate #4.47:** Lawrence Smith / Independent
- **Candidate #4.48:** Bill Kendrick / Independent

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

- **Candidate #4.49:** Fred Stein / Independent
- **Candidate #4.50:** Jerry Cole / Independent
-
- **Contest #5:**

Title of Office	Lieutenant-Governor
District of Office	Statewide
Partisanship	Partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	1
Maximum Write-in Votes Allowed	1

- **Candidate #5.1:** Chris Norberg / Blue
- **Candidate #5.2:** Anthony Parks / Yellow
- **Candidate #5.3:** Luis Garcia / Purple
- **Candidate #5.4:** Charles Qualey / Orange
- **Candidate #5.5:** George Hovis / Pink
- **Candidate #5.6:** Burt Zirkle / Gold
- **Candidate #5.7:** Brenda Davis / Gray
- **Candidate #5.8:** Edward Freeman / Aqua
- **Candidate #5.9:** Paul Swan / Brown

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

- **Contest #6:**

Title of Office	Registrar of Deeds
District of Office	Countywide
Partisanship	Partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	1
Maximum Write-in Votes Allowed	1

- **Candidate #6.1:** Laila Shamsi / Yellow

- **Contest #7:**

Title of Office	State Senator
District of Office	31st District
Partisanship	Partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	1
Maximum Write-in Votes Allowed	1

- **Candidate #7.1:** Edward Shiplett / Blue
- **Candidate #7.2:** Marty Talarico / Yellow

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

- **Contest #8:**

Title of Office	State Assemblyman
District of Office	54th District
Partisanship	Partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	1
Maximum Write-in Votes Allowed	1

- **Candidate #8.1:** Andrea Solis / Blue
- **Candidate #8.2:** Amos Keller / Yellow

- **Contest #9:**

Title of Office	County Commissioners
District of Office	Countywide
Partisanship	Partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	5
Maximum Write-in Votes Allowed	5

- **Candidate #9.1:** Camille Argent / Blue
- **Candidate #9.2:** Chloe Witherspoon / Blue
- **Candidate #9.3:** Clayton Bainbridge / Blue
- **Candidate #9.4:** Amanda Marracini / Yellow
- **Candidate #9.5:** Charlene Hennessey / Yellow
- **Candidate #9.6:** Eric Savoy / Yellow
- **Candidate #9.7:** Sheila Moskowitz / Purple
- **Candidate #9.8:** Mary Tawa / Purple
- **Candidate #9.9:** Damian Rangel / Purple
- **Candidate #9.10:** Valarie Altman / Orange
- **Candidate #9.11:** Helen Moore / Orange
- **Candidate #9.12:** John White / Orange
- **Candidate #9.13:** Joe Lee / Pink
- **Candidate #9.14:** Joe Barry / Pink

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

- **Candidate #9.15** Martin Schreiner / Gray

- **Contest #10:**

Title of Office	Court of Appeals Judge
District of Office	Statewide, 4th seat
Partisanship	Non-partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	1
Maximum Write-in Votes Allowed	1

- **Candidate #10.1:** Michael Marchesani

- **Contest #11:**

Title of Office	Water Commissioners
District of Office	City of Springfield
Partisanship	Partisan
Minimum Votes Allowed	0
Maximum Votes Allowed	2
Maximum Write-in Votes Allowed	2

- **Candidate #11.1:** Orville White / Blue
- **Candidate #11.2:** Gregory Seldon / Yellow

- **Contest #12:**

Title of Office	City Council
District of Office	City of Springfield
Partisanship	Partisan

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

Minimum Votes Allowed	0
Maximum Votes Allowed	4
Maximum Write-in Votes Allowed	4

- **Candidate #12.1:** Harvey Eagle / Blue
- **Candidate #12.2:** Randall Rupp / Blue
- **Candidate #12.3:** Carroll Shry / Blue
- **Candidate #12.4:** Beverly Barker / Yellow
- **Candidate #12.5:** Donald Davis / Yellow
- **Candidate #12.6:** Hugh Smith / Yellow
- **Candidate #12.7:** Reid Feister / Yellow

- **Retention Question #1:**

Wording of Question	Retain Robert Demergue as Chief Justice of the Supreme Court?
---------------------	----------------------------------------------------------------------

- **Retention Question #2:**

Wording of Question	Retain Elmer Hull as Associate Justice of the Supreme Court?
---------------------	---------------------------------------------------------------------

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

- **Referendum #1:**

Title of proposition	PROPOSED CONSTITUTIONAL AMENDMENT C
Wording of proposition	<p>Shall there be amendments to the State constitution intended to have the collective effect of ensuring the separation of governmental power among the three branches of state government: the legislative branch, the executive branch and the judicial branch?</p> <p>a. Article III, Section 6 of the Constitution shall be amended to read as follows:</p> <p>Section 6. Holding of offices under other governments. - Senators and representatives not to hold other appointed offices under state government. --No person holding any office under the government of the United States, or of any other state or country, shall act as a general officer or as a member of the general assembly, unless at the time of taking such engagement that person shall have resigned the office under such government; and if any general officer, senator, representative, or judge shall, after election and engagement, accept any appointment under any other government, the office under this shall be immediately vacated; but this restriction shall not apply to any person appointed to take deposition or acknowledgement of deeds, or other legal instruments, by the authority of any other state or country.</p> <p>No senator or representative shall, during the time for which he or she was elected, be appointed to any state office, board, commission or other state or quasi-public entity exercising executive power under the laws of this state, and no person holding any executive office or serving as a member of any board, commission or other state or quasi-public entity exercising executive power under the laws of this state shall be a member of the senate or the house of representatives during his or her continuance in such office.</p> <p>b. Article V of the Constitution shall be amended to read as follows: The powers of the government shall be distributed into three (3) separate and distinct departments: the legislative, the executive and the judicial.</p> <p>c. Article VI, Section 10 of the Constitution shall be deleted in its entirety.</p> <p>d. Article IX, Section 5 of the Constitution shall be amended to read as follows:</p> <p>Section 5. Powers of appointment.- The governor shall, by and with the advice and consent of the senate, appoint all officers of the state whose appointment is not herein otherwise provided for and all members of any board, commission or other state or quasi-public entity which exercises executive power under the laws of this state; but the general assembly may by law vest the appointment of such inferior officers, as they deem proper, in the governor, or within their respective departments in the other general officers, the judiciary or in the heads of departments.</p>

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

- **Referendum #2:**

Title of proposition	PROPOSED CONSTITUTIONAL AMENDMENT D
Wording of proposition	Shall there be an amendment to the State constitution concerning recovery of damages relating to construction of real property improvements, and, in connection therewith, prohibiting laws that limit or impair a property owner's right to recover damages caused by a failure to construct an improvement in a good and workmanlike manner; defining "good and workmanlike manner" to include construction that is suitable for its intended purposes; and permitting exceptions for laws that limit punitive damages, afford governmental immunity, or impose time limits of specified minimum lengths on filing lawsuits?

- **Referendum #3:**

Title of proposition	PROPOSED CONSTITUTIONAL AMENDMENT H
Wording of proposition	<p>Shall there be an amendment to the State constitution allowing the State legislature to enact laws limiting the amount of damages for noneconomic loss that could be awarded for injury or death caused by a health care provider? "Noneconomic loss" generally includes, but is not limited to, losses such as pain and suffering, inconvenience, mental anguish, loss of capacity for enjoyment of life, loss of consortium, and other losses the claimant is entitled to recover as damages under general law.</p> <p>This amendment will not in any way affect the recovery of damages for economic loss under State law. "Economic loss" generally includes, but is not limited to, monetary losses such as past and future medical expenses, loss of past and future earnings, loss of use of property, costs of repair or replacement, the economic value of domestic services, loss of employment or business opportunities. This amendment will not in any way affect the recovery of any additional damages known under State law as exemplary or punitive damages, which are damages allowed by law to punish a defendant and to deter persons from engaging in similar conduct in the future.</p>

- **Referendum #4:**

Title of proposition	PROPOSED CONSTITUTIONAL AMENDMENT K
Wording of	Shall there be an amendment to the State constitution authorizing

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

proposition	<p>Madison and Fromwit Counties to hold referenda on whether to authorize slot machines in existing, licensed parimutual facilities (thoroughbred and harness racing, greyhound racing, and jai alai) that have conducted live racing or games in that county during each of the last two calendar years before effective date of this amendment? The Legislature may tax slot machine revenues, and any such taxes must supplement public education funding statewide. Requires implementing legislation.</p> <p>This amendment alone has no fiscal impact on government. If slot machines are authorized in Madison or Fromwit counties, governmental costs associated with additional gambling will increase by an unknown amount and local sales tax-related revenues will be reduced by \$5 million to \$8 million annually. If the Legislature also chooses to tax slot machine revenues, state tax revenues from Madison and Fromwit counties combined would range from \$200 million to \$500 million annually.</p>
-------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

- **Referendum #5**

Title of proposition	BALLOT MEASURE 101: Open Primaries
Wording of proposition	Requires primary elections where voters may vote for any state or federal candidate regardless of party registration of voter or candidate. The two primary-election candidates receiving most votes for an office, whether they are candidates with no party or members of same or different party, would be listed on general election ballot. Exempts presidential nominations. Fiscal Impact: No significant net fiscal effect on state and local governments.

- **Referendum #6:**

Title of proposition	BALLOT MEASURE 106: Limits on Private Enforcement of Unfair Business Competition Laws
Wording of proposition	Allows individual or class action "unfair business" lawsuits only if actual loss suffered; only government officials may enforce these laws on public's behalf. Fiscal Impact: Unknown state fiscal impact depending on whether the measure increases or decreases court workload and the extent to which diverted funds are replaced. Unknown potential costs to local governments, depending on the extent to which diverted funds are replaced.

End of logical specification for Test Ballot Specification.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Appendix B

Instructions for Participants

In our mock election, we will be using fake names for candidates and colors for political party names. For example, you might see this:

Joe Jones/Yellow Party

Any similarity between names of candidates and real people or colors and real parties is purely coincidental.

Please attempt to vote exactly as described on the following pages

Once you start, we will not be able to help you.

Please do the best you can. If you are stuck and cannot continue, inform the administrator.

Thank you.

Usability Performance Benchmarks for the VVSG

For President and Vice President of the United States, vote for
Adam Cramer and Greg Vuocolo

For Senator, vote for
David Platt

For Congress, vote for
Brad Schott

For Governor, vote for
Cathy Steele

Do not cast a vote for
Lieutenant Governor

For Registrar of Deeds, write in a vote for
Christopher Christopher

For State Senator, vote for
Edward Shiplett

For State Assemblyman, vote for
Amos Keller

For County Commissioners, vote for the following candidates:
Camille Argent
Mary Tawa
Joe Barry

and enter write in votes for:
Dorothy Johns
Charles Blank

For Court of Appeals Judge, vote for
Michael Marchesani

For Water Commissioner, vote for
Orville White
Gregory Seldon

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

For City Council, vote for the following candidates:

Randall Rupp

Carroll Shry

Donald Davis

For Chief Justice of the Supreme Court

Vote to keep Robert Demergue in office

For the question of retaining Justice of the Supreme Court Elmer Hull

Do not vote

For Proposed Constitutional Amendment C

Vote for this amendment

For Proposed Constitutional Amendment D

Vote for this amendment

For Proposed Constitutional Amendment H

Vote against this amendment

For Proposed Constitutional Amendment K

Vote against this amendment

For Ballot Measure 101: Open Primaries

Do not vote

For Ballot Measure 106: Limits on Private Enforcement of Unfair Business Competition

Laws

Vote for the measure

Cast your ballot

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Appendix C

Frequently Asked Questions about This Research

How does one determine whether a voting system is sufficiently usable in actual elections based on using laboratory tests as opposed to using a real-world environment? Lab testing attempts to, as much as possible, mimic real-world voting environments, but at the same time there is so much diversity in election procedures across all voting locations in the U.S. that it is not possible to exactly duplicate or use all real-world voting environments for lab testing. Consistency of the testing is the focus here. The test shows that a specific set of people can vote accurately and can cast their ballots successfully on the specific ballot used in testing. The benchmark value used for this determination is an agreement of an acceptable level based on the specific ballot and test.

How can vendors be assured that their systems will be tested consistently, regardless of which test participants or labs are used? It will be demonstrated that each lab will be able to repeat the test and obtain the same data prior to having the labs conduct conformance tests on vendor equipment. Safeguards will be in place in the test to ensure that each test using a specific set of test participants is also a valid instance of the test protocol.

How can testing today's systems be used to set benchmarks for future systems? In general, benchmark development based on current systems can lag behind advances in technology. But, because benchmarks based on usability performance are technology independent, they can be used for identifying differences in the levels of usability of new systems. The current benchmarks were chosen such that several of the systems used for this research would have difficulty meeting them. In the future, the benchmarks can be adjusted upwards and made more stringent. Vendors and researchers can also use the test methodology on systems in development to measure usability improvements.

Is more research planned? Yes. The tests will be repeated in two other regions of the US and with other variations in the test participants, and future research will continue to address these benchmarks. Future tests will also use accessible voting systems and test participants with disabilities as described in the VVSG.

Can the results of this research predict the performance of a specific voting system in actual elections? Only indirectly. Real-world voting environments, ballots, poll workers, and voting procedures vary a great deal. The results of this research conducted in a laboratory setting, with a standard test ballot, and with specific sets of test participants cannot be used to predict performance of voting systems in actual environments. Like gas mileage ratings, your "mileage may vary" depending on the actual environment in which the voting system is deployed. The TGDC's primary goal

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.

Usability Performance Benchmarks for the VVSG

with this research is to provide a way to ensure that voting systems conforming to the VVSG have good usability.

This paper has been prepared at the direction of the HFP subcommittee and does not necessarily represent any policy positions of NIST or the TGDC.