Statement of Sharon Laskowski, NIST to the TDGC at the 9/22/04 Hearing:

Good afternoon and thank you for the opportunity to address the committee concerning usability testing of voting systems. I am speaking as a computer scientist at the National Institute of Standards and Technology. I am also the manager of the Visualization and Usability Group in the Information Technology Lab at NIST where we investigate user-centered evaluation methods and metrics for interactive systems and have helped to develop usability standards.

I am the primary author of the NIST Human Factors report written for the Help America Vote Act, which is entitled "Improving the Usability and Accessibility of Voting Systems and Products". The report was delivered to Congress in May of 2004 and can be found at vote.nist.gov and www.eac.gov. This report contains descriptions of most research and standards pertaining to usability and accessibility of voting systems as well as ten recommendations; some more difficult to implement than others, such as the development of performance based usability standards and tests for conformance to those standards.

Staff from my group, including myself, also lead the NIST Industry USability Reporting Project and one of the products of this international collaboration with top industry usability labs and usability professionals has been the Common Industry Format for Usability Test Reports which is now an ANSI/INCITS standard and soon to be approved by the ISO, the International Standards Organization. It captures the industry best practice of summative usability testing of software which is relevant to usability testing of voting systems as well.

I am here today to discuss what it means to test for usability and accessibility and how that relates to creating and testing to a usability performance-based standard. This type of testing is the only way to ensure high degrees of usability and the only way to incorporate usability reliably, and without bias toward a particular design or product, into voting system standards.

Let me give you some background about the current standards and testing program before I discuss this in more detail.

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

This report led the U.S. Congress to direct the Federal Elections Commission (FEC) to work with NIST to conduct a study of the feasibility of developing national standards for

voting systems. Following release of the 1982 report, limited funds were appropriated to begin the multi-year effort. Thirteen meetings and five years later, with the help of about 130 different policy and technical officials, the FEC instituted the 1990 Voluntary Voting System Standards (VSS).

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

The VSS themselves were substantially updated and issued again in 2002. This release included functional requirements to improve accessibility by individuals with disabilities. An advisory section was included as guidance to improve user interface and ballot design. There were no specific qualification test criteria developed for this section; hence no formal conformance tests are associated with the guidance.

As we have heard for these three days and in many other venues, the current standards and testing process need improvement. And, in particular, standards for usability and accessibility need to be created. (Parenthetically, it should be noted that independent test labs run conformance tests for standards in many other domains and there are accreditation processes that can be put into place to ensure a high quality testing process. There are other models. Conformance test suites are sometimes executed by the vendor (self-testing), or by a potential purchaser. But it has become common practice for a third party, such as an accredited laboratory to perform the testing.)

There are few usability standards in place except for some design guidance such as font size, clearly labeled fields, and privacy considerations. These provide no guarantee that a voter can cast a vote as intended quickly and correctly. And, in some cases, the standards are ambiguous and therefore difficult to test.

Even more troubling is that we have very little data about the levels of usability of voting systems. Until very recently there has been little applied research from the human factors and usability fields specifically on voting systems. (Examples of such research include the work of Susan King Roth, the NSF-funded project led by Paul Herrnson, and that of Ted Selker as part of the CalTech-MIT Voting Technology Project.) Accessibility has been addressed by generic design standards that intended to remove barriers to access, but usability by persons with disabilities has not been addressed by research. In fact, we know very little about users' experiences with voting systems including those people with disabilities. This suggests a need to focus efforts on building a foundation of applied research for voting systems and voting products to support the development of usability standards. Until this is done, there is little basis upon which to include many detailed design specifications.

So, the issue is how to develop voting system standards for usability that are performance-based. These would contain benchmarks, based on performance that would guarantee a certain level of usability. The questions, then, are: What metrics should be used for performance? How do we measure them? How do we decide on the benchmarks, the levels of performance? There are 3 metrics that are the "classic" ones for usability, as described in ISO 9241-11.

- Effectiveness (e.g., voter votes for intended candidate, no errors)
- Efficiency (e.g., voter completes voting task within reasonable amount of time and effort)
- Satisfaction (e.g., voter's experience is not stressful, voter is confident)

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

Further, these are the same metrics that are used in summative usability testing in industry.

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

The development of a good test suite can often involve more effort than the formulation of the standard itself.

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

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

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

Though formative or diagnostic tests are valuable tools in the design process, they do not guarantee that the final product is usable as measured by the metrics described earlier

(efficiency, effectiveness, and satisfaction) since they are used during the design process, not on a final product. Even tests that are conducted on a final product design are generally not conducted in a way that would allow the results to be generalized to the intended populations (i.e., the participants of the study may or may not be appropriately extrapolated to a majority of all actual users). This is particularly true for voting system products since the range of users required for such a test would make this type of testing cost prohibitive to most vendors. In addition, there are currently no defined standards for usability metrics that vendors could use as benchmarks for their testing. For these reasons, it is clear that vendor testing of the product, while valuable, is a separate issue from certifying that the end product is usable. Usability qualification and certification testing is necessary and it will require the establishment of both objective usability test procedures and pass/fail criteria or benchmarks.

A valid, reliable, repeatable, and reproducible process for usability testing of voting products against agreed-upon usability benchmarks is more complex than typical summative usability testing because of the diversity of the voter population and the need for very low error rates. In particular, there must be a careful definition of the metrics, such as what counts as an error, how to measure error rate, time on task, etc., by which systems are to be measured. The benchmarks for error rates should be restricted to usability problems leading to partial failure, and usability problems leading to total failure. Since we are dealing with outcomes, usability problems prior to success need not be specifically included, but would be represented in the time on task measure from testing. Note that while excessive time required does not lead to failure, it is still unacceptable.

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

Research needs to be conducted to determine: the nature of errors possible during a voting process (this includes voter errors and poll worker errors), and the level (rate) of these errors (both the current levels for existing products and recommendations for "acceptable" levels of each error type). Once this information is available, I recommend that a set of repeatable and reproducible processes be defined and that each voting product is tested using these test processes and usability test benchmarks. This would include the definition of all test procedures, the data collection required, the data analysis approach, participant screening and selection procedures, and reporting requirements.

The key issues we face, and the research that needs to be done, is to define a valid and reliable test that can generalize to the voting experience in terms of both a “standard range” of subgroups of the voting population and a "standard range" of ballots.

While field testing for usability should be done with the specific ballot intended for a specific election, certification cannot be done using a single ballot but rather requires the range of ballots previously agreed upon. These ballots will need to be representative of some fairly large percentage of all ballots the machine will be expected to support in the field in terms of length and complexity.

Initial experiments can help determine what performance benchmarks are realistic, and one would hope that these can be improved over time as vendors become more experienced with user centered design and usability testing. As for the pass/fail vs. product comparison, conformance tests are typically pass/fail for many standards. Also, at this point there is not enough research to determine whether variations in performance that surpass a benchmark would be statistically meaningful and lead to valid, reliable comparisons.

Though the ITAs would likely have the responsibility to conduct these tests, the nature and format of the testing would likely require additional personnel with qualifications to conduct this type of testing.

Any such test process would have to work across all populations and for different types of elections. This range has to be reasonable but it will never be complete. The limitations of the certification process would be the equivalent of the statements made in gas mileage ratings for cars. There's an agreement on how it will be measured and reported, but there's also a careful note that tells you that your mileage may vary since the mileage ratings given are under test conditions. In other words, if there is an election in which a machine that has been shown to be acceptable will be used, but the ballot design so complex that it was never considered in conformance testing, your mileage may vary from that reported by the testing authority.

Finally, although the issues that I have outlined for usability standards and testing will require some research and experimentation, I believe that the issues can be addressed to a great extent in a relatively short time frame, say 1- 2 years given appropriate funding.