Comments for Panel on Usability Testing Voting Systems
EAC Technical Guidelines Development Committee
NIST
September 22, 2004

Frederick Conrad
University of Michigan

Thank you for the opportunity to address members of the EAC/TGDC, thank you to NIST for hosting these hearings and to Whitney Quesenbery for organizing today's panels. I am an Associate Research Scientist at the University of Michigan, Institute for Social Research where I teach courses in survey methodology. My background is in cognitive psychology and current research interests include usability of web surveys and, more recently, usability of electronic voting systems.

I am here on behalf of an interdisciplinary research team (four political scientists, one computer scientist and one psychologist) working on a project funded by the National Science Foundation's Digital Government program on usability of electronic voting systems. The team members are Paul Herrnson and Ben Bederson from the University of Maryland, Dick Niemi from Rochester University, Mike Hanmer from Georgetown University and Mike Traugott and myself from the University of Michigan.

Our project involves a complementary approach of field and laboratory usability testing but I will focus my remarks on our laboratory experiences. I'm happy to discuss the field tests (carried out in shopping malls and senior centers) during the question and answer session. However, I'd now like to tell you about our recent laboratory test so I can refer back to it in answering the committee's questions.

In late July and early August, 42 members of the Ann Arbor, MI community visited our laboratory at the University of Michigan and voted for the same fictional candidates and ballot questions on six electronic voting machines: four touch screen systems, one machine with hardware buttons and a dial for navigation and selection, and one optical scan machine. The ballot was either an office block ballot (organized by the race) or a straight party ballot (organized by party and allowing the user to vote entirely for Democrats or Republicans with a single or small number of button presses or other actions).

The users were instructed to vote in all races except one where we explicitly instructed them to abstain. In addition we instructed them to change a vote in one race and write-in the candidate in another. The voters' interactions were video taped and after voting on each machine they completed a satisfaction questionnaire about that machine. When they completed the entire laboratory task users were asked to answer an additional questionnaire about their background and some general beliefs about electronic voting. I

will not be talking about the results of the study –we have barely begun to analyze the data – but instead the methods we used and methodological lessons we have learned.

Turning now to the first question … and I'll spend most of my time answering the first one…

1. *How should we conduct usability testing of voting systems, given their unique requirements?*

Actually, we don't think voting systems have such unique requirements when it comes to evaluating their usability so I will answer this question by describing three ways in which usability testing of voting interfaces ought to resemble usability testing of other kinds of user interfaces: (1) the usability measures, (2) the ability to compare across interface designs, and (3) scenario based testing.

1.1 Measures

Just as in the testing of most user interfaces the key measures are going to concern users' speed, accuracy and satisfaction in completing (or not completing) the voting task.

Speed is a relatively straightforward measure. The main concern is at what level of granularity do we measure speed – time to submit the overall ballot, time to record votes in particular races, time to change a vote, time to write-in a vote, etc.? By video taping all interactions, it is possible to measure any of these and more.

Accuracy (and it's complement—error) is a little more complicated. Here the notion of *slips* versus *mistakes* (Norman, 1981) is relevant. A slip is an unintended action like a stray finger movement grazing a touch screen while a mistake is an intended but erroneous action like deliberately pressing the "cast votes" button after choosing a candidate for the first race because the user believes that each vote must be submitted separately.

Perhaps the most serious mistake one can make when it comes to voting is to select a candidate other than the one the voter intends to select. Such a mistake could occur, for example, because the angle of the screen leads a voter to believe she is touching the region for one candidate when in fact she is touching the region for another. To record the voter's intent, we provided them with an information booklet about the fictional candidates and ballot questions at the start of the testing session. We asked them to circle the candidates for whom they intended to vote or, in some cases, for whom we intended them to vote. To the extent that their votes do not match the circled candidates and choices on ballot questions, they have made a mistake (in our technical sense). We will determine who they actually voted for by viewing the video tapes and examining the ballot images.

Errors of omission include undervoting and the related phenomenon of roll-off (no votes cast beyond a certain point in the ballot). This is difficult to measure in the laboratory because users are paid to complete the ballot. We can, however, observe how they react

to feedback from the voting interface indicating they have undervoted. By instructing them not to vote in particular races we assure that we can test each system's mechanism for dealing with undervotes. The feedback is implemented in different ways in different designs, e.g. in some cases it is immediate and local and in others it is provided in a review screen.

So errors need to be classified in some way and tallied.

User satisfaction. Some measure of the users' subjective experience is essential because users who are performing well may not complete the task or may not return to vote in the next election unless the experience is relatively pleasing and free of frustration. To assess this, it is typical to administer a satisfaction questionnaire immediately after the session. There are many commercially (and freely) available, general purpose user satisfaction questionnaires for which the items have been demonstrated to provide reliable measures but which are not tailored to any specific software product. We tailored some of the questions to make them more sensitive to voters' – as opposed to generic users' – experience. The trade-off is that the measurement properties of these items are unknown. We administered this right after the voter cast her ballot on each of the six machines that we tested.

1.2 Value of Comparisons across interfaces: Evaluating the usability of a single design – whether a voting system or a web site – is almost sure to be worthwhile and informative. Evaluating more than one design and comparing results is even more so. Consider vote changing. Some systems require deselecting an already selected vote by repeating whatever action was used to select it in the first place, e.g. pressing a checked-check box. Others allow users to simply press the field for the new candidate. One can imagine that the first design would be problematic for voters who are not familiar with today's interface conventions; alternatively one can imagine that the latter might be confusing for people who are. If only one design were tested, this important comparison would not be possible. This is sound practice in general and makes sense for the evaluation voting system usability as well.

1.3 Importance of Scenario based testing. We scripted the task for our users by requiring them to vote for whichever candidates they indicated on the voter information booklet and by requiring them to write-in a vote, change a vote and not vote in one race. This was the same for all six machines. As a result we can compare the identical task across different machines so that differences cannot be due to differences in task. Scenarios also allow us to stress features of the design that we believe up front are likely to be problematic.

Having just argued that usability testing methods for voting systems are not unique, one way in which they do pose special requirements is that they must be usable by all citizens. If ever there was a technology for which universal usability is essential it is this technology. I will return to this shortly but my main point is that if a voting system is to be considered usable not only must those with sensory and motor impairments be able to successfully cast their vote but so must those without technological expertise be able to

vote with confidence. The digital divide – at least the version related to those who can figure out how to use unusable systems and those who cannot – must be bridged for this technology.

A related idiosyncrasy of evaluating voting system usability is that even infrequent problems can be of great importance in close elections so the typical practice of testing 4 to 6 users could fail to detect such problems. We tested 42 users in the laboratory and even this may not be enough to identify infrequent problems.

An advantage of testing a relatively large number of laboratory users is that it makes it possible to detect some quantitative differences statistically. We believe that quantitative data is essential (though not sufficient by itself) for measuring voting system usability.

2. *What role can usability testing play in the certification process, or to provide inputs to the certification process?*

One could establish absolute levels on any of the types of measures we have been discussing (time, accuracy and satisfaction) and treat these as criteria for certification. For example, usability tests can establish a range times across machines in which most people should be able to accurately complete a particular voting task, e.g. voting for all races on a particular ballot. For a machine to be certified, it would need to be usable by test voters within these ranges.

3. *How do we ensure that the participants in usability testing represent the full spectrum of voters?*

In laboratory tests this is hard to do. The approach we adopted was to radically over-recruit respondents who we believed were likely to experience usability problems. As I indicated earlier, it was our intuition that those unfamiliar with the conventions of graphical user interfaces would be far more likely to have problems than more computer literate users. Thus of our 42 users, 30 had limited computer experience. By "limited experience" we mean less than once every six months. In many cases these users indicated they had never used a computer. Many used email somewhat often but did nothing else. This also meant that our users were disproportionately older.

4. *What research needs to be done to provide input to human factors and accessibility standards for voting systems?*

We have barely begun to explore usability of voting systems and so know relatively little about problems that may be unique to the topic[1]. Some areas where we badly need research are:

[1] There may well be unique usability problems with voting systems even if the way we test the systems and detect those problems is not unique.

- Usability of multi-modal/multi-media interfaces, e.g. usability of synthesized speech output and spoken user input, particularly when coordinated with other modes and media, e.g. speech output combined with tactile input, or visual output combined with spoken input.

- Usability across different ages. Do the elderly, who generally have smaller working memory capacity than younger people – and turn out to vote in large numbers – have more trouble with paging designs than full face or, for example, zooming designs in which it is possible to see the whole ballot and less necessary to remember previous displays?

- Usability for election officials: setting-up, closing-down, transmitting data, providing help to voters. Is it reasonable to require the design to be usable without occasional help? If not, what amount of help is optimal, i.e. speeds the process for any one voter without slowing the process for everyone else by overwhelming the available resources?

- Usability of printed ballot receipts in conjunction with touch screens (or other electronic displays): If paper records for verification purposes become common place, their use will requires the voter to split his attention between the two devices – the one displaying the printed ballot receipt and the one on which he has indicated his choices – since it is on the latter device that he must verify the former (at least in current designs) by indicating that what is printed conforms to what he registered on the user interface. What kinds of problems are caused by this split attention?

- Public perceptions of voting machines and their usability: are citizens deterred from going to their polling place on Election Day because they believe the machines will be too hard to use?

References

Norman, D. (1981) Categorization of action slips. *Psychological Review, 88* 1-15