JUS
JOURNAL OF USABILITY STUDIES

Vol. 2, Issue 2, February 2007, pp. 96-98

Comments on: Selker, Rosenzweig, and Pandolfo (2006). "A Methodology for Testing Voting Systems," JUS, Volume 2, Issue 1, 7-21.

Whitney Quesenbery
John Cugini
Dana Chisnell
Bill Killam
Ginny Redish

In the article, "A Methodology for Testing Voting Systems" (JUS, November 2006, pp7-21), Selker, Rosenzweig, and Pandolfo discuss their methodology for usability testing of voting systems. With so much at stake in the usability of our ballots and voting systems, we can only applaud any research in this field. There is little history of research in this area, so discussions of test protocols are especially valuable. Unfortunately, although this article sets out to compare "the relative merit in realistic versus lab style experiments for testing voting technology," it falls short of this goal. If their point is that real-world testing is important because real election environments add burdens that are not present in lab settings, this conclusion is not supported by any of the work described.

More importantly, there is a surprising gap in this article on methodology: a lack of any discussion of the impact of the research goals for a study of voting systems on the best protocol for that research. The article is written as though there is just one "good" methodology, and our job is to find it. But, a methodology that is good for one purpose might be less than ideal for another.

A usability test, specifically a test of a voting system, might have two different kinds of goals. It might be:

- A test designed for either diagnosis or measurement. That is, are we trying to understand how the system works, what errors were made and why, with an aim to improve the design. Or, are we trying to accurately measure the performance of the system. A diagnostic test might have less controlled experimental conditions and rely more on qualitative analysis and expert interpretation.
- A test focusing on either the voting process or on the machine itself. The former would focus not only on how participants completed their tasks, but on the effect of different environmental conditions (such as lighting, noise, instructions, or interactions with poll workers). A test of the machine (or paper ballot) would use a more controlled environment to reduce these variables.

Given the importance of understanding the research goal, the gaps in the details of the three usability studies described in the article make it impossible to assess their claims. Without that information, the test descriptions are interesting, but do not make a substantive contribution to the development of methodologies for testing voting systems.

It is also surprising that in a political environment such as voting, the article fails to reveal the purpose of the studies, or who commissioned them. Were they simply research projects at MIT, or were they constructed to test a specific hypothesis? Even overlooking these omissions, however, the article skips over some critical issues in considering the relative merits of different usability testing protocols.

In both of the studies that used simulated polling places, a very low percentage of participants produced usable data. In the *New York Reading Disabilities* study, “These conditions led to many problems, and only data from 41 subjects were able to be collected.” (p9-10). This is out of a pool of 97 participants. In the *Arlington Voter Verification Study*, “…35 out of 48 subject of the participant data being useful.” (p15). That is less than 50% and 70%, respectively. This is particularly troubling in a test of voting systems, where all voters should be able to complete basic voting tasks. It also poses a serious methodological question for anyone planning to use a simulated polling place for other research, but there is no explanation for why so many participants were dropped from the results, or any discussion of possible solutions for this problem.

The Arlington study seems to have been plagued with problems, including prototypes that did not work well, “poll worker confusion” (p15) about the experimental process, and participants who “purposely decided not to follow the instructions and wouldn’t vote for the candidates pre-assigned on the voting card.” Despite this, the authors conclude that “best practice…includes…recognizable candidates” (p19). There is no discussion of the alternative approach discussed by other researchers in this field (including University of Michigan, ACCURATE and NIST researchers): using realistic, but fictitious, names and ballot questions, or of the differences between fictitious names and party names.

Problems with compliance to the test instructions is particularly important for usability tests of voting systems. Observation of user actions is particularly difficult, as these systems typically cannot be instrumented for recording. This leaves only direct

observation, or positioning a camera in the voting booth, but either of these destroys the simulation of the polling place in a country with a secret ballot. Unfortunately, the authors are silent on how they addressed this central question in testing voting systems.

The authors' conclusions are confusing. The weight of evidence seems to suggest that their naturalistic test settings create procedural problems, and offer little advantage over a laboratory setting:

- "Both (of the) studies raised question of whether the effort of creating an ecologically valid voting experience improves or weakens the data compared to normative testing of the equipment in a laboratory setting." (p16)
- "The study of the voting methodology in semi-naturalistic settings does give important and rich results that do point to important research directions." (p17)
- "Comparing the results …further validates the procedure of testing in laboratory settings."

There is no discussion of the specific aspects of the realistic setting that the authors found valuable, except to say that "it also allows for discovery of usability issues….with the process and environment of the act of voting" (p17). This may be true, but given the wide variation in voting procedures across the country, it might be useful only for the jurisdiction the test procedures are based on. And, that may—or may not—be a goal for the research.

The article claims that testing in the real world is better than testing in a lab setting. This claim is hard to argue with, but not adequately supported in the paper. Even as desirable as real world testing is, the issues associated with full real world testing for voting are so great as to make it nearly impossible to produce valid, reliable, and reproducible results. We think "realistic" (as opposed to real) testing is the only thing that is practical.

Reference

Everett, Sarah, Michael Byrne, and Kristen Greene, "Measuring the Usability of Paper Ballots: Efficiency, Effectiveness, and Satisfaction," Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting. Santa Monica, CA: Human Factors and Ergonomics Society, 2006. Available at http://chil.rice.edu/research/pdf/EverettByrneG_06.pdf