A Comparison of Usability Between Voting Methods

Kristen K. Greene, Michael D. Byrne, and Sarah P. Everett
Department of Psychology
Rice University, MS-25
Houston, TX 77005 USA
{kgreene, byrne, petersos}@rice.edu

ABSTRACT

In order to know if new electronic voting systems truly represent a gain in usability, it is necessary to have information about the usability of more traditional voting methods. The usability of three different voting methods was evaluated using the metrics recommended by the National Institute of Standards and Technology: efficiency, effectiveness, and satisfaction. Participants voted with two different types of paper-based ballots (one open-response ballot and one bubble ballot) and one mechanical lever machine. No significant differences in voting completion times or error rates were found between voting methods. However, large differences in satisfaction ratings indicated that participants were most satisfied with the bubble ballot and least satisfied with the lever machine. These data are an important step in addressing the need for baseline usability measures for existing voting systems.

Author Keywords
Usability, voting, paper ballots, mechanical lever machine

INTRODUCTION

Usability is a critical issue for voting systems; even if a voting system existed that was perfectly secure, accurate, reliable, and easily auditable, it could still fail to serve the needs of the voters and the democratic process. Voting systems must also be usable by all voters. A truly usable voting technology should be a walk-up-and-use system, enabling even first time voters to cast their votes successfully. It should be accessible for those voters with disabilities, and it should not disenfranchise particular groups of voters. The Help America Vote Act (HAVA) of 2002 was a landmark piece of legislation in terms of calling public attention to issues such as these.

From a human factors and usability standpoint, voting is an arena where research could and should have great practical impact. Voting is a challenging usability problem for two main reasons: characteristics of the user population and characteristics of the task itself. The user population is extremely diverse, representing a wide range of ages and varying greatly in education, socioeconomic status, familiarity with technology, etc. In addition to this extremely diverse voter population, there are subpopulations of users—such as those those who are visually impaired or physically disabled, or those who do not speak and/or read English—for whom voting has historically been particularly challenging. Also important from a usability standpoint are the numerous first time voters and the relatively infrequent nature of the task—even the most zealous voter will not have the opportunity to vote more than a few times per year. This makes it especially crucial that voting systems are usable. One of the promises of electronic voting technology is that it may help address this issue.

However, there is currently no baseline usability information for existing voting technologies to which new systems can be compared. The goal of the current study is to begin addressing this need for baseline data. It is the second in a series of experiments we are conducting that compares the usability of multiple voting technologies: paper ballots, optical scan ballots, mechanical lever machines, punchcard voting systems, and DREs.

The three usability measures used in the current study are those metrics recommended by the National Institute of Standards and Technology (NIST) (Laskowswki, et al., 2004): efficiency, effectiveness, and satisfaction. These are also suggested by the International Organization for Standardization (ISO 9241-11, 1998).

Endorsement by NIST is important, because unlike other usability domains, decisions about acceptability are made by governmental standards bodies.

Efficiency is defined as the relationship between the level of effectiveness achieved and the amount of resources expended, and is usually measured by time on task. Efficiency is an objective usability metric that measures whether a user's goal was achieved without expending an inordinate amount of resources (Industry Usability Reporting Project, 2001). Did voting take an acceptable amount of time? Did voting take a reasonable amount of effort?

Effectiveness is the second objective usability metric recommended by NIST, and is defined as the relationship between the goal of using the system and the accuracy and completeness with which this goal is achieved (Industry Usability Reporting Project, 2001). In the context of voting, accuracy means that a vote is cast for the candidate the voter intended to vote for, without error. Completeness refers to a vote actually being finished and officially cast. Effectiveness is usually measured by collecting error rates, but may also be measured by completion rates and number of assists. If a voter asks a poll worker for help, that would be considered an assist.

Satisfaction is the lone subjective usability metric recommended by NIST, and is defined as the user's subjective response to working with the system (Industry Usability Reporting Project, 2001). Was the voter satisfied with his/her voting experience? Was the voter confident that his/her vote was recorded? Satisfaction can be measured via an external, standardized instrument, usually in the form of a Likert scale questionnaire. The System Usability Scale, (SUS), (Brooke, 1996) is a common standardized instrument that has been used and verified in a multitude of domains.

Everett, Byrne, & Greene (2006) used measures of efficiency, effectiveness, and satisfaction to compare the usability of three different types of paper ballots. The current study expands upon this previous research by using the same three measures to assess the usability of mechanical lever machines, in addition to measuring usability for two of the three previously-studied paper ballots.

The three types of paper ballots evaluated by Everett, Byrne, & Greene (2006) were bubble ballots, arrow ballots, and open response ballots (see Figure 1 for examples of bubble and open response ballots). Every participant voted on all three ballot types. Each participant saw either realistic candidate names or fictional candidate names, but never both. Realistic names were those of people who had either announced their intention to run for the office in question, or who seemed most likely to run based on current politics. Fictional names were produced by a random name generator. Participants were also provided with one of three different types of information. They were given a slate that told them exactly who to vote for, given a voter guide, or were not given any information about candidates at all.

The three dependent measures used in the previous study were ballot completion time, error rates, and satisfaction. Ballot completion time (efficiency) measured how long it took the participant to complete his or her vote. No significant differences were found between ballot types (bubble, arrow, or open response) or candidate types (realistic or fictional) on the ballot completion time measure. Significant differences were observed between information types; participants given a voter guide took reliably more time to complete their ballots than those who received a slate or those who did not receive any information about candidates.

Errors (our measure of effectiveness) were recorded when a participant marked an incorrect response in the slate condition or when there was a discrepancy between their responses on the three ballot types. Overall, error rates were low but not negligible, with participants making errors on slightly less than 4% of the races.

Satisfaction was measured by the participant's responses to the SUS questionnaire. The bubble ballot produced significantly higher SUS scores than the other two ballots, while the open response and arrow ballot were not significantly different from one another.

Given that we found no significant differences on ballot completion times between realistic or fictional candidate names, for the current study only fictional names were used. This decision was made to prevent advertising campaigns and media coverage from differentially affecting data from one study to the next. In addition, the continued use of realistic candidate names would have quickly rendered the study materials obsolete, making comparisons between studies with new materials difficult. Using fictitious names circumvents both these issues for the current study and future research. Since no significant differences were found between ballot completion times for the three paper ballot types in our previous study, the decision was made to only examine two of those three for the current study. Since the current study evaluated an additional voting technology, mechanical lever machines, we chose to retain two of the three previously-studied paper ballots in order to be able to define errors by majority agreement.

METHOD

Participants

Thirty-six people participated in the current study, 13 of whom were Rice University undergraduate students and 23 of whom were recruited from the larger Houston population. There were 21 female and 15 male participants ranging in age from 18 to 56. The average age was 28.61 (SD = 10.48). All had normal or corrected-to- normal vision and were US citizens fluent in English. On average, participants had voted in 3.11 national elections, ranging from zero to 12. Seventeen of the 36 participants had only voted nationally once before, and another four participants had no prior national voting experience. On average, participants had voted in 7.06 other types of elections (local, school, etc.). Rice students received credit towards a course requirement for participation and all other participants were compensated $25 for the one-hour study.

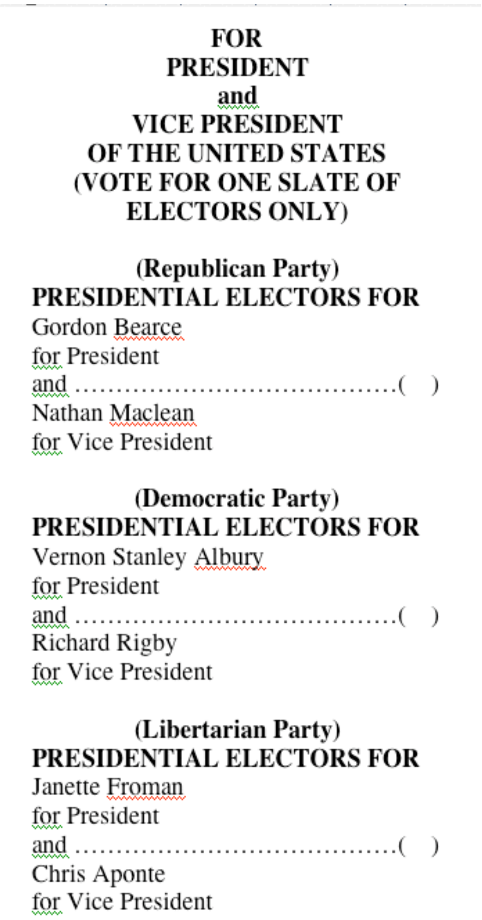
Design

A mixed design with one between-subjects variable and one within-subjects variable was used. The between-subjects manipulation was information condition. This meant that each participant received only one of three types of information: slate, voter guide, or no information. A slate was a paper that listed exactly who the participant was to vote for, as well as whether to vote yes/no on the propositions. The voter guide used was based on those produced by the League of Women Voters. It was left completely up to the participant whether she/he chose to actually read and use the voter guide. In the control condition, participants were not given any information about the candidates or propositions. The within-subjects variable was voting method. This meant that every participant voted using each of the three following voting methods: open response ballot, bubble ballot, and mechanical lever machine. The order in which participants voted was counterbalanced, so that each voting method was used first, second, and third an equal number of times.

The three dependent variables measured in the study were voting completion time, errors, and satisfaction. Voting completion time was measured in seconds and indicated how long it took a participant to complete a ballot. Errors were recorded when a participant marked an incorrect response in the slate condition or when there was a discrepancy between their responses on the three voting methods. Satisfaction was measured by the participant's responses to the SUS.

Materials and Procedure

The same fictional candidate names used by Everett, Byrne, & Greene (2006) were used for each of the three voting methods in this study. The first voting method was the open response ballot (Figure 1A). The second voting method was the bubble ballot (Figure 1B).



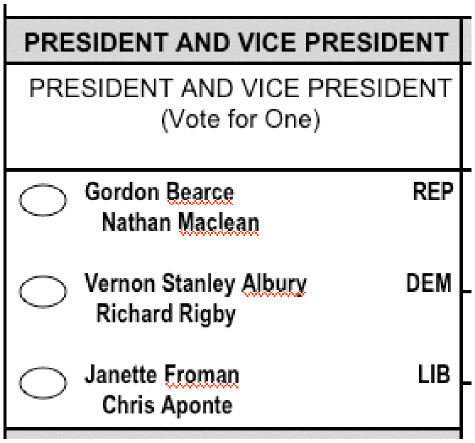
FOR
PRESIDENT
and
VICE PRESIDENT
OF THE UNITED STATES
(VOTE FOR ONE SLATE OF
ELECTORS ONLY)

(Republican Party)
PRESIDENTIAL ELECTORS FOR
Gordon Bearce
for President
and ....( )
Nathan Maclean
for Vice President

(Democratic Party)
PRESIDENTIAL ELECTORS FOR
Vernon Stanley Albury
for President
and ....( )
Richard Rigby
for Vice President

(Libertarian Party)
PRESIDENTIAL ELECTORS FOR
Janette Froman
for President
and ....( )
Chris Aponte
for Vice President

Figure 1A. Open response ballot



PRESIDENT AND VICE PRESIDENT

PRESIDENT AND VICE PRESIDENT
(Vote for One)

○ Gordon Bearce
Nathan Maclean REP

○ Vernon Stanley Albury
Richard Rigby DEM

○ Janette Froman
Chris Aponte LIB

Figure 1B. Bubble ballot

The third voting method was the mechanical lever machine (Figure 2). The lever machines used in this study were purchased at auction from Victoria County, TX.

Each voting method used the races and propositions used by Everett, Byrne, & Greene (2006). Candidate names were produced by a publicly-available random name generator (http://www.kleimo.com/random/name.cfm). The 21 offices ranged from national races to county races, and the six propositions were real propositions that had previously been voted on in other counties/states, and that could easily be realistic issues for future elections in Harris County, TX, the county in which this study was conducted.

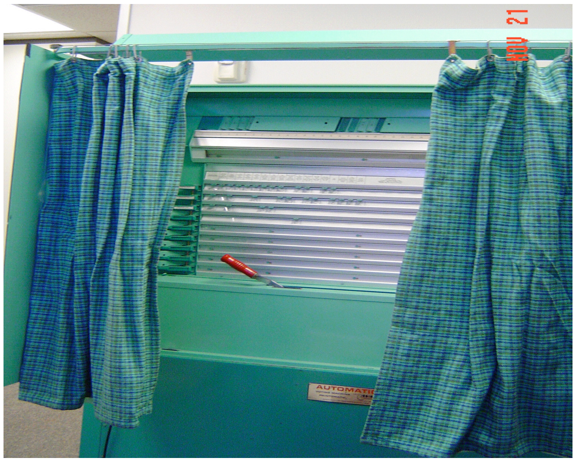


Figure 2. Mechanical lever machine

Participants first went through the informed consent procedure and read the instructions. Those in the control condition were allowed to begin voting immediately, those in the slate condition were given their slate before beginning to vote, and those in the voter guide condition were given ample opportunity to read over the voter guide. Participants stood throughout the voting process, either at a voting station when using the paper ballots, or behind the curtains of the lever machine when using that voting method.

After voting, participants completed a survey, then were informed about the purpose of the study and paid. The survey included both general demographic questions and questions more specific to voting itself. Voting specific questions asked about people's previous voting experience, as well as their opinions on the exact voting methods used in the study. Also part of the survey was a separate SUS for each of the three voting methods. The SUS is a Likert scale, which has people answer questions using a scale of one to five, ranging from strongly disagree to strongly agree. The SUS includes questions such as “I thought the system was easy to use” and “I felt very confident using the system.” Responses to ten such questions are then combined to yield a single number; the higher the final number, the higher the degree of satisfaction a person experienced (Brooke, 1996).

RESULTS

Efficiency

Figure 3 displays the voting completion times for the three voting methods evaluated in this study. No significant differences in response times were found between the three voting methods, F(2, 66) = 0.24, *p* = 0.78, nor was there a significant effect of information type on response times, F(2, 33) = 2.01, *p* = 0.15. Finally, the interaction of voting method and information type was not significant, F(4, 66) = 0.29, *p* = 0.89.

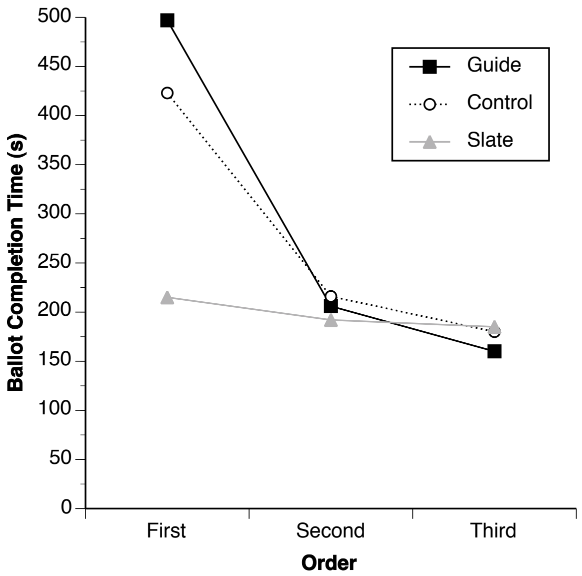


Figure 3. Voting completion times by voting order and information condition

However, people were reliably slower with the first voting method they used; there was a significant effect of the order in which participants used the three voting methods, F(2, 66) = 27.81, *p* < 0.001. There was also a significant interaction of voting method order and information type, F(4, 66) = 5.64, *p* = 0.001. This effect is driven entirely by the first ballot completed by the participants. Participants in the “slate” condition were significantly faster on their first ballot than participants in the other two conditions. However, participants in the “guide” and control conditions did not reliably differ on their first ballot, and all three conditions were statistically indistinguishable on their second and third ballots.

Effectiveness

We computed error rates two ways: by race and by ballot. “By race” rates were calculated by dividing the number of errors made by the total number of opportunities for error. There were 27 races (21 offices plus six propositions) on each of the three voting methods. Therefore each participant faced 81 total opportunities for error. Error rates for each ballot type are presented in Table 1. Differences between ballot types on error rate were not significant, F(2, 66) = 0.43, *p* = .66.

	Open	Bubble	Lever	Overall
Error Rate	.012	.009	.007	.009

Table 1. Error rates by voting method

Similarly, participants did not make significantly more errors depending on what type of information they were given; there was no significant effect of type of information given, F (2, 35) = 0.13, *p* = .88. This is important because it indicates that the error rate seen in this study is not simply a memory problem whereby participants were failing to remember from one ballot to the next who they voted for.

Another way to consider error rates is by ballot. A ballot either does or does not contain one or more errors. Frequencies by ballot type are presented in Table 2. A chi-square test of association revealed no reliable effect of voting method ($X^2(2) = 3.42$, *p* = .18), that is, voting method did not appear to affect the frequency with which ballots containing errors were generated. However, a surprising 17 of the 108 ballots collected contained at least one error; this means nearly 16% of the ballots contained at least one error.

Errors			
	None	At least 1	Total
Bubble	32	4	36
Open	27	9	36
Lever	32	4	36
Total	91	17	108

Table 2. Ballot error frequencies by voting method

Satisfaction

Participants were most satisfied when using the bubble ballot and least satisfied when using the lever machine; in the SUS scores (Figure 4). This effect was statistically reliable, F(2, 66) = 9.37, $p < 0.001$. Posthocs reveal that this difference was driven almost entirely by the difference between the bubble ballot and the lever machine, which were significantly different. However, differences between the open ballot and the other ballots were not.

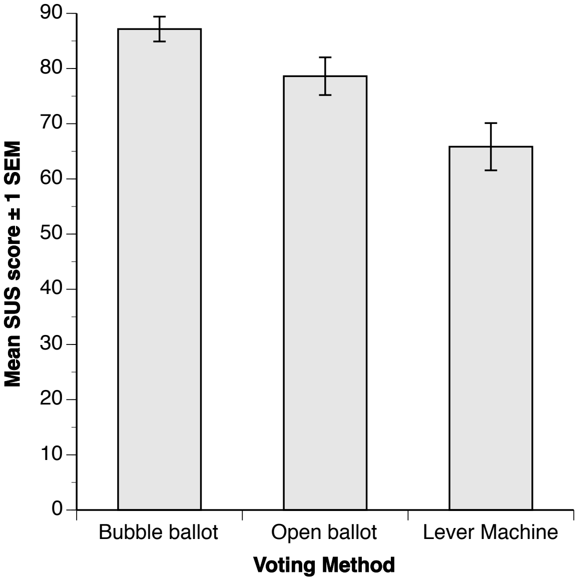


Figure 4. SUS scores

DISCUSSION

In terms of efficiency, as measured by time taken to vote, the three voting methods used in this study seemed approximately equivalent. Effectiveness, however, is clearly a concern. While the overall rate of .009 error per race appears low, this overall effect on ballots is quite striking. (In fact, if errors were statistically independent of one another, an error rate of .009 per race predicts that in a 27-race ballot, just over 20% of the ballots should contain at least one error.) The substantial error rate seen indicates possible room for improvement in the effectiveness of paper ballots and lever machines.

Despite these similarities between voting methods on the usability metrics of efficiency and effectiveness, sharp differences emerged when examining the usability metric of satisfaction. When Everett, Byrne, & Greene (2006) found that bubble ballots elicited the highest satisfaction ratings, it was unclear whether that was an artifact of the population from which they sampled. It may not have been surprising that they found Rice University undergraduates to be most satisfied when using bubble ballots, since they so closely resemble standardized tests—something with which they are very familiar as students. It is interesting that the bubble ballot continued to receive the highest satisfaction ratings in the current study, one in which a more diverse and representative sample was used.

FUTURE DIRECTIONS

While the sample used in the current study was an improvement over previous work, it should be broadened even further in several important respects. As the oldest participant in the current study was only 56 years old, a crucial next step will be to sample from a much wider age range. Also, by testing participants who do not read/ speak English, one may see interesting cultural differences emerge. Study materials are being translated into both Spanish and Chinese and future studies will recruit from these subpopulations. Similarly, the range of voting methods should be broadened to include punch card voting, a system which was one common but which is currently on the decline. From a usability perspective, how bad (or good) are punch card systems?

Even with a more representative sample, many unanswered questions remain. For example, ballot layouts and location/clarity of instructions will always be a potential issue. In our study, the lever machines present all candidates from a particular political party in a single horizontal row, which makes it easier for people who vote straight-party ticket to complete and check their vote. Is this a visual display characteristic that is worth emulating in the layout of paper ballots or DREs as well? While DREs and lever machines can both be programmed to allow straight-party ticket voting by a single lever pull or on-screen selection, paper ballots do not have such an option. How would adding a straight-party ticket choice affect the efficiency and effectiveness of paper ballots? Would there be ramifications for absentee ballots?

The rapid adoption of DREs raises another important series of questions. Are the DREs actually easier to use than their traditional counterparts? What are the implications of voter verified paper audit trails (VVPATs) for usability? Clearly, such paper trails must cost voters time if they are accurately verifying them, but how much time, and how accurate are voters when examining such audit trails?

These kinds of questions only scratch the surface of the total space of human factors issues in elections, issues which will need to be addressed to ensure confidence in any voting system, electronic or otherwise.

ACKNOWLEDGMENTS

We would like to thank Phil Kortum and Stephen Goggin for comments on and assistance with this research. This research was supported by the National Science Foundation under grant #CNS-0524211 (the ACCURATE center). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of the NSF, the U.S. Government, or any other organization.

REFERENCES

Brooke, J. (1996) SUS: a “quick and dirty” usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & A. L. McClelland (eds.) *Usability Evaluation in Industry*. London:Taylor and Francis.

Everett, S. P., Byrne, M. D., & Greene, K. K. (2006). Measuring the usability of paper ballots: Efficiency, effectiveness, and satisfaction. To appear in *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

Industry Usability Reporting Project. (2001). Common industry format for usability test reports (ANSI/INCITS 354-2001).

International Committee for Information Technology Standards. ISO 9241-11. (1998). Ergonomic requirements for office work with visual display terminals (VDT)s part 11. Guidance on usability.

Laskowski, S. J., Autry, M., Cugini, J., Killam, W., & Yen, J. (2004). Improving the usability and accessibility of voting systems and products. NIST Special Publication 500-256.

Reference:

Greene, K. K., Byrne, M. D., & Everett, S. P. (2006). A comparison of usability between voting methods. To appear in *Proceedings of the 2006 USENIX/ACCURATE Electronic Voting Technology Workshop*. Vancouver, BC, Canada.