MEASURING THE USABILITY OF PAPER BALLOTS: EFFICIENCY, EFFECTIVENESS, AND SATISFACTION

Sarah P. Everett, Michael D. Byrne, and Kristen K. Greene
Department of Psychology, Rice University
Houston, TX

The Help America Vote Act (HAVA) of 2002 secured funding for improvements to election administration. Improvements include upgrading older voting systems to meet new guidelines. To determine whether the new voting systems are improvements over existing voting systems, information is needed on the usability of the older, traditional systems. This study was designed as a first step in addressing the need for usability data on existing voting systems. Three traditional paper ballots were empirically evaluated to collect baseline data that can later be compared to newer, electronic voting systems. Usability was evaluated using the thee International Organization for Standardization (ISO) metrics suggested by the National Institute of Standards and Technology (NIST): effectiveness, efficiency, and satisfaction. All three ballot types (bubble, arrow, and open response) produced reasonable levels of efficiency. The three ballot types did not produce different levels of effectiveness, but the overall error rate was higher than would be expected. On satisfaction, voters were clearly more satisfied with their experience with the bubble ballot.

The Help America Vote Act (HAVA) of 2002 secured funding for improvements to election administration. Improvements include upgrading older voting systems to meet new guidelines. Older systems such as lever machines and punchcard ballots are widely being replaced with electronic voting systems. These new systems are touted as better for multiple reasons. For example, they can easily accommodate voters requesting foreign language ballots, and, due to their potential for improved accessibility, their adoption has been supported by disability groups such as National Federation of the Blind.

Although there have been recent attempts to provide guidelines for voting systems and standards for testing them, there is still much work to be done. The U.S. Election Assistance Commission (EAC) has released the Voluntary Voting System Guidelines (VVSG) as a suggested set of standards to ensure that voting systems provide acceptable levels of usability and security. To determine whether the new voting systems are improvements, information is needed on the usability of the older, traditional systems. Empirical data on usability of these systems is thus required.

There are many things which make usability for voting systems challenging. Among these is that the user population is extremely diverse, including users with physical or visual impairments, those who neither read nor speak English, and tremendous diversity in education and socioeconomic status. Also, voters receive little or no training on using the voting method and voting is a fairly infrequently performed task.

A report issued by the National Institute of Standards and Technology (NIST) (Laskowski, et al., 2004) suggested using three International Organization for Standardization (ISO) usability metrics to evaluate voting systems: effectiveness, efficiency, and satisfaction. Unlike many other human factors domains, usability of election equipment may indeed be mandated by government standards, and NIST is the federal body which sets such standards.

First, the effectiveness of systems needs to be measured. An effective system is one in which the voter's intent is correctly represented. The voter must be able to cast a vote for the candidate for whom they intend to vote without making mistakes. The infamous "butterfly ballot" in the 2000 election in Florida is a case where effectiveness was a serious issue.

The second recommended usability metric is efficiency; on usable systems, voting should take reasonable time and effort. Voting is a unique situation because of the sheer number of people who must use a system in a single day. When the efficiency of a system is inadequate, lines at the polls can quickly reach unreasonable lengths.

User satisfaction is the third usability metric recommended by the 2004 NIST report and is the only subjective measure of the voting experience. Even if a voting system produces votes for intended candidates and allows for efficient voting behavior, it will not be a success unless the voters themselves consider it to be so. The voting experience should not be stressful and voters should not feel confusion. Voters should be confident that their vote was cast accurately and will be counted.

The current study is the first in a series of experiments examining the usability of voting systems and comparing usability across technologies. The design of ballots and voting systems affects voter's feelings about their ability to cast a vote and this can influence how they feel about the legitimacy of an election (Niemi & Herrnson, 2003). While these effects are known, the evidence needed to support recommendations for new voting system designs or ballot redesigns is lacking. While various researchers have studied how the physical layout of a ballot or voting technology can affect issues such as residual

votes, overvoting, and unrecorded votes (Ansolabehere & Stewart, 2005; Herron & Sekhon, 2003; Kimball & Kropf, 2005), there has not been much of an attempt to compare overall usability of disparate voting systems. Benchmark data on the usability performance of voting methods is needed as a standard for the conformance testing of new voting technologies (Laskowski & Quesenbery, 2004) and to be able to compare various voting systems.

This study was designed to be a starting point in addressing the need for usability data on existing voting systems. Traditional paper ballots, both those which have been used and those still in use, were empirically evaluated. These evaluations were completed to generate baseline data that can later be compared to newer, electronic voting systems.

METHOD

Participants

Forty-two Rice University undergraduates participated in this study. There were 27 female and 15 male participants ranging in age from 18 to 22. All were U.S. citizens and fluent in English. Of the 42 participants, 23 had previously voted in a national election. 36 participants received credit towards a course requirement and 8 received $10 for their participation.

Design

There were three independent variables used in the study: ballot type, candidate type, and the amount of information given. Ballot type was a within-subjects variable with three levels: bubble, arrow, and open response. Candidate type was a between-subjects variable and was whether the participant saw the names of real or fictional candidates. The other between-subjects variable, amount of information given, had three levels: slate, guide, and no guide. In the "slate" conditions, participants were given a list of candidates and propositions to vote for. Participants in this condition did not decide for themselves how to vote; they were instructed to vote exactly how the slate indicated. In the "guide" condition participants received a voter guide describing the races and made their own decisions. Participants decided how carefully they read the voter guide. If they typically use a voter guide in real-life voting situation, they could read the guide thoroughly. If they do not usually look at a voter guide, they could simply set it aside. Participants in this condition made their own choices about for which candidates or propositions to vote. Finally, in the "no guide" condition participants were given no information about the candidates. These participants were simply told to vote however they wished. Because information type was a between-subjects variable, each participant only received one type of information; they received only one of the following: a slate, a voter guide, or no information.

The three dependent variables in the study were ballot completion time, errors, and satisfaction. These corresponded to the NIST's suggestions for the usability metrics of effectiveness, efficiency, and satisfaction. Ballot completion time measured how long it took the participant to complete each ballot and was measured in seconds. 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 ballots. Satisfaction was measured by the participant's responses to a subjective usability scale.

Materials

Ballots. The first ballot was the "bubble" ballot. It was a optical scan-style ballot that required participants to fill in the bubble next the their choice (much like standardized test forms). The "arrow" ballot was also inspired by optical scan ballots on which participants indicated their choice by drawing a line to complete a broken arrow pointing to their choice. The third type of ballot was an open-response ballot (inspired by ballots used in Mississippi). This ballot contained a pair of parentheses next to each choice, but did not provide information about how to indicate a choice. See Figure 1 for examples of each ballot type.

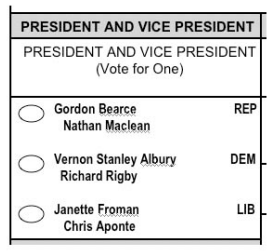

Figure 1A

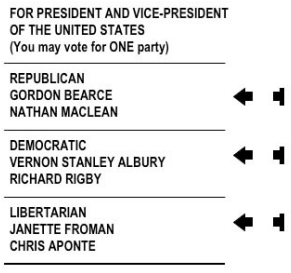

Figure 1B

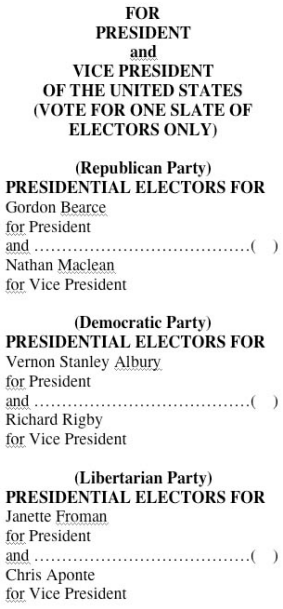

Figure 1C

Figure 1. Examples of the ballots used in study. 1A. The bubble ballot, 1B. The arrow ballot, 1C. The open-response ballot.

All experimental ballots were based on true sample ballots in the 2004 national election and closely resembled the examples. Because the presentation order of the ballots was counterbalanced, some participants saw bubble ballot first, followed by the arrow or open response, while others saw the

arrow ballot first followed by the open response or bubble ballot, etc.

As mentioned above, there were two types of candidates: realistic and fictional. The realistic ballot contained the names of real people who could potentially run for office in the next national or local election. The fictional names were produced by a random name generator (http://www.kleimo.com/random/name.cfm) and the obscurity factor was set to 15 to reflect the diversity present on many ballots. The gender of the fictional candidates matched that of the realistic candidates.

Because candidate type was a between-subjects variable, each participant saw either realistic or fictional candidates throughout the experiment. Disclaimers were put on the guides, instructions, and debriefing that reminded participants that the materials and information in the guide were strictly for research purposes and may not accurately reflect the views of any real person.

Each ballot contained 21 races and 6 propositions. The offices ranged from that of president at the national level to tax collector at the county level. The six propositions used were ones that had not been previously voted on in a real election in the vicinity of the study, but had relevance to the area and could perhaps be proposed in a future election. The same propositions were used on the on the realistic and fictional ballots.

Voter Guides. The design of the voter guides was based on those distributed by the League of Women Voters of the Houston Area. For the realistic candidates, the guides presented the true views of the candidates as much as possible. For the fictional candidates, their positions on various issues were constructed to resemble those of the realistic candidates.

Slate. Participants in the slate condition received a paper with each race in the election. This sheet had the name of the person for whom the participant was to vote and whether they should vote yes or no for the propositions.

There were three types of slates used in the study. The mixed slate contained an approximately equal mix of Republicans and Democrats. The primarily Republican or Democratic slate instructed participants to vote in 85% of the races for the Republican or Democratic candidate. All slates instructed participants to vote yes for four propositions and no for two. The assignment of the slates was counterbalanced.

Survey. The survey packet that participants completed included questions about demographics, previous voting experience, and the ballots used in the study. This survey packet also included a System Usability Scale (SUS) scale for each of the three ballot types. The SUS is a Likert scale and is comprised of 10 questions related to various aspects of satisfaction. The ratings for each item and combined to produce a single satisfaction score. Scores on this scale range from 0 to 100 (Brooke, 1996) with higher scores indicating higher perceived usability.

Procedure

Participants stood at a voting station while completing their ballots. To make the situation as realistic as possible, the stations used were the actual ones used in previous national elections, which had been donated to the experimenters. After signing an informed consent form and reading the instructions for the experiment, participants were handed the first of their three ballots. The presentation order of the three ballot types was counterbalanced.

After completing the each ballot, participants returned it to the experimenters and were handed the next ballot. They were reminded be as consistent as possible in the candidate choices. After returning the second ballot, participants were handed and completed their third and final ballot. If the participant was in either the guide or control condition, the experimenters completed a short exit interview with the participant when the third ballot was returned. This simply required the participants to verbally explain for which candidates and proposition they had voted. This was done to obtain the participant's intent, without the intermediary of a ballot. In the slate condition, the exit interview was not conducted because the candidate or proposition of the participant's intent could be read from their slate. Finally, in all conditions, participants filled out the survey packet before receiving the debriefing.

RESULTS

Ballot Completion Times

Table 1 displays the ballot completion times for the three ballot types evaluated in the study. There were no significant differences between ballot types on the ballot completion time measure, $F(2,72) = 0.57$, $p = .57$.

Ballot Type	Mean	SD
Bubble	324.3	423.3
Arrow	284.8	518.9
Open	235.3	248.1

Table 1. Ballot completion time means and standard deviations by ballot type (in seconds).

While there was not an effect of ballot type, information and order did affect completion time; see Figure 2 the relevant graph. Both main effects of order, $F(1, 36) = 25.71$, $p < .001$, and information condition, $F(2, 36) = 9.99$, $p < .001$, were significant, as was the interaction, $F(2, 36) = 9.93$, $p < .001$. These effects were not surprising; participants took longer in the guide condition, participants took longer on their first ballot, and participants in the guide condition took an especially long

time on their first ballot. This effect was expected as many participants carefully read the guide to decide on candidates and propositions.

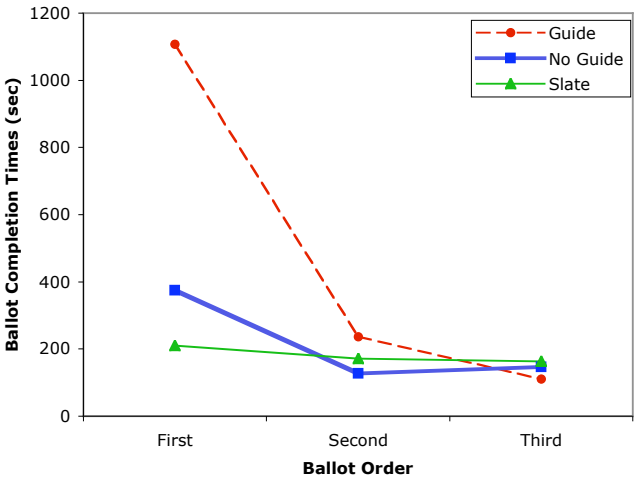

Figure 2. Ballot completion times by ballot order and information type.

There was not a significant difference between the realistic and fictional candidate conditions, $F(1, 36) = 0.08$, $p = 0.78$.

Errors

We computed error rate two ways: by race and by ballot. "By race" error rates were calculated by dividing the number of errors made by the number of opportunities for errors. There were 21 candidate races and 6 propositions on each ballot, and each participant voted on 3 ballots. This meant that there were 81 opportunities for errors. The error rates by ballot and overall are shown in Table 2. Differences between ballot types were not statistically significant, $F(2, 72) = 0.72$, $p = .49$

	Open	Bubble	Arrow	Overall
Error Rate	.009	.013	.009	.010

Table 2. The overall error rate and the error rate by ballot type.

Furthermore, there were no significant differences in error rates by candidate type, $F(1, 36) = 1.20$, $p = 0.28$), information condition, $F(2, 36) = 0.47$, $p = .63$, nor were there reliable interactions. This meant that the frequency of errors was not affected by the use of fictional candidates. More importantly, because there was not a significant difference between the error rates of those who received a slate of candidates for whom to vote and those that did not, the errors made by participants were unlikely to be a result of memory problems.

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 ($X^2(2) = 0.64$, $p = .73$), that is, voting method did not appear to affect the frequency with which ballots containing errors were generated. However, a surprising 14 of the 126 ballots collected contained at least one error; this is over 11% of the ballots.

Errors			
	None	**At least 1**	**Total**
Bubble	36	6	42
Open	38	4	42
Arrow	38	4	42
Total	112	14	126

Table 3. Error-containing ballots by ballot type.

SUS

In the SUS scores, there was a surprisingly large difference between ballot types; see Figure 3. This difference was statistically significant, $F(2, 82) = 9.52$, $p < .001$. Pairwise comparisons revealed that the bubble ballot produced significantly higher SUS scores than the other two ballots ($p = .001$ and $p < .001$). The open-response ballot and arrow ballot were not significantly different from each other ($p > 0.50$). There were no effects or interactions of the other independent variables on SUS scores.

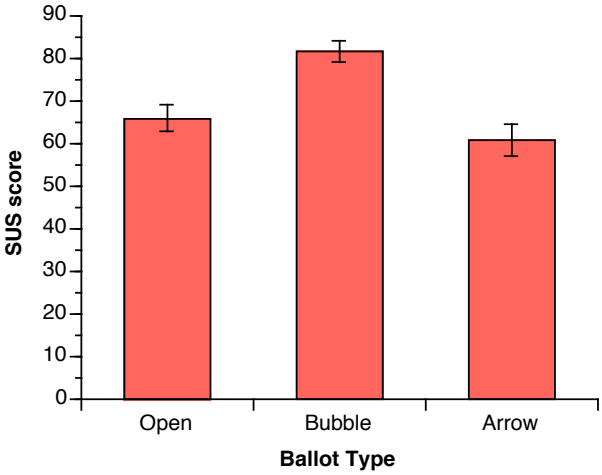

Figure 3. SUS means and SEM error bars for the three ballot types.

In addition to individual ballot preferences, survey responses also revealed that in previous voting experiences in real elections, 12% of participants had been unsure whether their vote was cast correctly or would be counted. A larger 26% had previously worried about figuring out how to use a ballot or voting technology to cast their vote. These numbers are surprisingly large for such a highly educated and technologically aware group.

DISCUSSION

All three ballot types produced reasonable levels of efficiency; that is, participants were able to complete their ballots in a reasonable amount of time (an average of under 5 minutes for a 21-race ballot).

Effectiveness, however, is clearly a concern. Whereas most studies of error in this domain rely on indirect measures such as overvote rates, our measure was much more direct. While the overall rate of 0.01 error per race (1%) appears low, the cumulative impact of this rate on longer ballots is quite striking. The fact that over 11% of the ballots contained at least one error is a clear cause for concern, with possible public policy implications such as procedures for handling of narrow margins of victory and recounts.

With respect to satisfaction, voters were clearly more satisfied with their experience with the bubble ballot. This was shown by the SUS scores for each ballot and was seconded by survey data that revealed a clear preference for the bubble ballot. While participants in the current study were most satisfied with their voting experience with the bubble ballot, it is important to keep in mind that these were current college students. The bubble ballot is very similar to standardized multiple-choice forms that are commonly used when administering exams. This means that the participants in this study were most likely quite familiar with this response format and this familiarity may have influenced their feelings of satisfaction when using the bubble ballot. However, what bodes well for future studies is that differences between multiple forms of paper ballot were discriminable; it seems likely that differences between entirely different forms of voting technology (e.g., lever machines and punch cards) would be even larger.

The survey data results are also compelling. The fact that such a high percentage of this sophisticated sample self-reported issues with previous voting experiences indicates there may be substantial subjective usability issues present in extant voting equipment.

LIMITATIONS

The population used in the current study is clearly not representative of the complete voting demographic. This particular population is highly homogenous on many demographic measures. However, they still had concerns with casting their vote accurately. If this population of bright and confident students had concerns about their ability to accurately cast their vote, a stronger effect might be expected in a less homogenous sample.

FUTURE WORK

Work including a larger, more representative sample is underway. It will also broaden to include more types of voting systems. Baseline usability information will need to be collected for other traditional technologies such as lever machines and punchcard ballots. Once this information has been collected, the usability of electronic voting systems can be compared to that of previous voting technologies and improvements or areas of concern can be discussed.

ACKNOWLEDGMENTS

We would like to thank the National Science Foundation for its support 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

Ansolabehere, S., & Stewart, C., III. (2005). Residual votes attributable to technology. *Journal of Politics, 67*(2), 365–389.

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.

Herron, M. C., & Sekhon, J. S. (2003). Overvoting and representation: an examination of overvoted presidential ballots in Broward and Miami-Dade counties. *Electoral Studies, 22,* 21–47.

Kimball, D. C., & Kropf, M. (2005). Ballot design and unrecorded votes on paper-based ballots. *Public Opinion Quarterly, 69*(4), 508–529.

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.

Laskowski, S. J., & Quesenbery, W. (2004). Putting people first: The importance of user-centered design and universal usability to voting systems, National Research Council, Committee on Electronic Voting, National Academies Press.

Niemi, R. G., & Herrnson, P. S. (2003). Beyond the butterfly: The complexity of U.S. ballots. *Perspectives on Politics, 1* (2), 317–326.

Reference: 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.