Electronic Voting Machines versus Traditional Methods: Improved Preference, Similar Performance

Sarah P. Everett*, Kristen K. Greene*, Michael D. Byrne*°, Dan S. Wallach°, Kyle Derr°, Daniel Sandler°, Ted Torous°
*Department of Psychology
°Department of Computer Science
Rice University
6100 Main St.
Houston, TX 77005 USA
{petersos, kgreene, dwallach, byrne, derrley, dsandler, ttorous}@rice.edu

ABSTRACT
In the 2006 U.S. election, it was estimated that over 66 million people would be voting on direct recording electronic (DRE) systems in 34% of the nation's counties [8]. Although these computer-based voting systems have been widely adopted, they have not been empirically proven to be more usable than their predecessors. The series of studies reported here compares usability data from a DRE with those from more traditional voting technologies (paper ballots, punch cards, and lever machines). Results indicate that there were little differences between the DRE and these older methods in efficiency or effectiveness. However, in terms of user satisfaction, the DRE was significantly better than the older methods. Paper ballots also perform well, but participants were much more satisfied with their experiences voting on the DRE. The disconnect between subjective and objective usability has potential policy ramifications.

Author Keywords
Voting, electronic voting, DRE, usability, preference

ACM Classification Keywords
H5.2 Information interfaces and presentation (e.g., HCI); User Interfaces

INTRODUCTION
In U.S. elections, voters are often presented with a multitude of races and candidates to consider. There can be national, state, and local races on the same ballot, and there may be one or many candidates running in each of these races. Voters are given the opportunity to choose one person in each of the races presented to them, but they are not required to vote in every race if they do not choose to do so. The ballots and voting methods can take many forms from paper-based to electronic.

The variety of voting methods and large numbers of candidates and races can lead to confusion and error. The problems in the 2000 U. S. Presidential election in Florida focused national attention on the need for usable voting systems. As the country became familiar with terms such as "butterfly ballot" and "hanging chads," many states decided to replace these systems to avoid such problems in future elections. The Help America Vote Act (HAVA) of 2002 provided funding for updating voting equipment and intended for states to replace their outdated voting methods with newer, more reliable systems. Because of this legislation and its requirement that election equipment be replaced by 2006, millions of dollars have been spent purchasing direct recording electronic (DRE) systems to replace older technologies.

Why Usability?
To ensure the integrity of elections, voters must be able to cast their votes as intended. Unintentional undervotes (i.e., not casting a vote in a race), overvotes (i.e., voting for more candidates in a race than is allowed), or votes for the wrong candidates can substantially impact elections, as evidenced in the 2000 election upset in Florida. Wand et al. [23] showed that the butterfly ballot used in Palm Beach County, Florida caused over 2,000 voters to vote for Pat Buchanan instead of Al Gore, tipping the election to George W. Bush. In another analysis of this election, Mebane [16] used ballot-level data to show that 50,000 votes intended for Bush or Gore were lost to overvotes. He claims that had technology been available to warn voters of overvotes, Gore would have won by more than 30,000 votes.

Voter confusion is especially problematic in elections because it is the voters themselves who must consider their voting experience to be a success. To have confidence in the outcome of an election, voters must believe that the record of their votes accurately captures their intent and that the final tally of the votes accurately includes all cast votes. Usability problems can also cause long lines at the polls. It

CHI 2008, April 5–10, 2008, Florence, Italy.

is important that voting not take an inordinate amount of time, otherwise many voters may be pragmatically unable to cast their vote. Finally, usability can affect future voter turnout. If voters have had a bad experience, believe their vote will not count, have to wait an extremely long time to vote, or worry about figuring out how to use a voting system, they may choose to abstain from voting in the future. The consequences of poor usability are magnified in the United States with its tendency to have a large number of issues presented to the voter at once, versus other countries where voters may only be asked to vote on a single issue.

Assessing Usability

To assess the usability of voting systems, the National Institute of Standards and Technology (NIST) recommends using the International Standards and Organization's (ISO 9241-11, 1998) [14] metrics of efficiency, effectiveness, and satisfaction [15]. As NIST is responsible for setting voting system testing standards, it is important for research on the usability of such systems to use these metrics.

Efficiency is an objective metric defined as the relationship between the level of effectiveness achieved and the amount of resources expended, and is usually measured by time on task. Effectiveness, also an objective metric, refers to the relationship between the goal of using the system and the accuracy and completeness with which this goal is achieved. Effectiveness is usually measured by collecting error rates, but may also be measured by completion rates and number of assists. The third usability metric recommended by NIST is satisfaction, defined as the user's subjective response to working with the system. Satisfaction can be measured via an external, standardized instrument, such as the System Usability Scale (SUS) [3], which has been used and verified in multiple domains [1].

While commercial DREs are what voters actually use to cast votes, they have certain limitations as a usability testing platform. They do not allow for easy modification by researchers so that the design space can be explored. Detailed information about user actions (e.g., timestamps) can be highly useful in usability assessment, but such information is not collected by commercial DREs, nor should it be, for privacy reasons. For these and other reasons, we chose to develop our own DRE, VoteBox.

Our VoteBox software platform supports a broad array of electronic voting research efforts; for example, it uses a pre-rendered user interface to minimize the complexity of the software stack that must be trusted to faithfully record the voter's intentions [26]. Additionally, it can be deployed in a networked configuration to provide fault-tolerance, tamper-evident audit logs, and an administrative interface that assists poll workers in correct operation of the voting equipment [21].

The version of VoteBox used in these studies comprises 8674 lines (4339 semicolons) of Java source code and runs on Windows, Mac OS X, and Linux systems. It is a DRE voting system that accepts user input via mouse; we are in the process of adding support for other input methods, including touch screens and hardware buttons.

The interactive ballot presented to the voter (shown in Figure 1) is similar to other DREs in common use: it presents a number of informational screens, followed by a number of contests, a review screen, a final confirmation screen, and a terminal screen. Contests consist of a number of choices that may be directly selected by the voter; the interface permits undervotes but prevents overvotes. The voter may move backward and forward among all but the last of these screens using on-screen controls. In addition to providing these basic voting functions, this version of VoteBox has been specifically modified to capture timing and other data necessary to support our studies.

Previous Research

Traditional measures of residual vote rate (overvotes and undervotes) are not hard to determine in real elections, but this does not include errors where the voter casts a vote for the wrong candidate, and not all undervotes are errors. In order to more directly measure error, researchers must conduct simulated elections where voter intent is clear and unambiguous. Previous work in this vein has examined baseline usability data for several traditional voting systems such as paper ballots, punch cards, and lever machines [5, 10, 12]. Paper ballots take many forms, but usually require the voter to use a pencil to make a mark next to the candidate of their choice on a pre-printed form containing many races. With the punch card system, the voter slips the punch card behind a booklet containing the races and candidates and uses a stylus to punch out the chads corresponding to the candidates of their choice. On a lever machine, voters are presented with all the races at one time with levers beneath each candidate. Voters indicate their choice by pulling the levers corresponding to the candidates for whom they would like to vote.

Overall, paper ballots, especially bubble ballots (so-named for the oval next to each candidate's name that voters must fill in to indicate their choice), seem to be the voting method that is the most usable for the greatest range of users. This is due in part to their error rate of about 1.5% [5]. While this number is certainly higher than we would like to accept in a task this important, it is lower than the error rates for both the punch cards and lever machines.

In the past few years, researchers have begun to look at the usability of DREs. Although they have not directly compared their findings on DREs to traditional forms of voting, Herrnson and colleagues [2, 6, 13, 22] have examined voters using commercially-available DREs. In more recent work (not yet been published), they found evidence of serious usability problems with DREs. In a large-scale field study, error rates (including residual votes and votes for the wrong candidate) on the presidential race were as high as 4.2% on one electronic voting system. Of

all voters in the study, 2.5% voted for the wrong candidate for president. Herrnson and colleagues point out that this is an especially serious type of error because not only does the preferred candidate lose a vote, but an opponent also gains a vote. This study also showed evidence that error rates were affected by voter demographics such as age, education, computer experience, race, and English as primary language.

These studies are important as they compare different types of DREs, and their results begin to shed light on the usability of electronic voting systems. What these studies do not tell us is how DREs compare directly to previous voting methods. The following studies attempt to answer this question.

EXPERIMENT 1

Experiment 1 was comprised of two nearly identical experiments, 1A and 1B. For both of these experiments, VoteBox was configured so that the basic model for navigating through the system was similar to most commercially-available DREs. Users (voters) were first presented with an instruction screen, followed by a series of screens that presented contests. Navigation between these screens was through "Next Page" and "Previous Page" buttons. At the end of the sequence, users saw a review screen which allowed them to double-check their choices, after which they could cast their ballot.

It should be noted that these experiments were part of a larger project that examined the effects of tampering with the content of the review screens. That issue is beyond the scope of the current paper; for additional information on that and more complete details on demographics, see [9].

Methods

Methods for Studies 1A and 1B were identical except where noted.

Participants

All participants were recruited through advertisements in local newspapers, and were paid $25 for their participation in either of the one-hour studies. In both studies, participants were fairly diverse in terms of education, ethnicity, and income. A 10-point Likert scale was used to assess self-rated computer expertise, with 1 representing a computer novice and 10 an expert.

The 66 participants in Experiment 1A ranged in age from 18-76, with a mean age of 43.1 years (*SD* = 15.5). 31 were male and 35 were female. 88% of the participants had previous voting experience. Participants were fairly comfortable with computers, rating themselves on average at 5.7 out of 10 (*SD* = 2.5).

There were 101 participants in Experiment 1B. There was an even gender split in the study with 51 male and 50 female participants. The age range of participants was 20-75, with a mean of 40.8 (*SD* = 14.8). 92% of participants had previous voting experience, and the average self-rating of computer expertise was 5.8 (*SD* = 2.4).

Design

Experiment 1A was a 2 x 3 between-subjects design with random assignment to the six conditions. The 2-level variable was information condition: half the participants were in an undirected condition in which they were offered a voter's guide and allowed to make their own selections. The second condition was a directed condition in which participants were given a piece of paper (a "slate") that specified which candidates they should choose. Since a choice was indicated for every race on the ballot, this condition will be referred to as "directed with no roll-off." ("Roll-off" refers to the tendency of voters to abstain from voting for races further down the ballot.) The second between-subjects experimental factor was the type of additional voting method used. Each participant was randomly assigned to vote with one of a paper ballot, a punch card, or a lever machine along with the DRE.

Experiment 1B was a 3 x 3 where a third level of information condition was added: some participants were given a paper instructing for whom they should vote, but this slate directed them to not select a candidate in some races. This was done to simulate real roll-off. The locations of skipped races on the slate were selected such that they represented real roll-off rates (i.e., the races further down the ballot, the more local races, were more likely to be skipped [18]). Thus, in 1B, the information condition variable had 3 levels: undirected, directed with no roll-off, and directed with moderate roll-off.

Materials

The DRE used was the VoteBox system created by the authors. An example of a candidate race screen generated with this system is depicted in Figure 1.

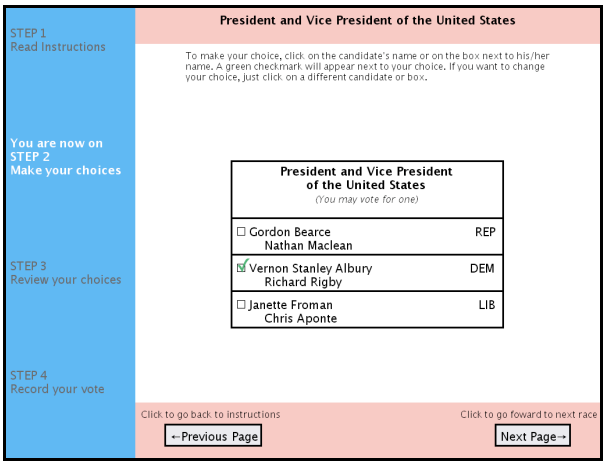

Figure 1. Screenshot of a candidate race.

The paper ballots, punch cards, and mechanical lever machines in these studies were all used in previous research [5]. There were 27 contests on each ballot, comprised of 21

offices and six propositions. Candidate names and voter guides were fictitious, taken from prior work [5, 10, 12] (this work has also shown the use of fictitious names did not affect the usability of different voting systems).

The SUS [3] was used to assess subjective user experience with each voting method. Additionally, a study-specific survey packet was created, containing questions about general demographics, previous voting experience, proficiency with computers, etc.

Procedure

Instructions given to participants differed depending on information condition. Participants in the directed conditions were told to vote as their slate instructed. Participants in the undirected condition were given a chance to study a voters' guide which described the candidates and their positions and were instructed to vote consistently, i.e. choose the same candidates each of the two times they voted. Participants cast two ballots, one with each of voting technologies to which they were assigned (the DRE and one of the additional methods; order of presentation was counterbalanced across subjects). The SUS was administered directly after each ballot was completed. If the participants were in the undirected condition, a brief exit interview was conducted after voting to obtain the voter's intent. All participants completed a survey packet.

Although punch card ballots and optical-scan ballots like our paper ballot are typically scored by machine, they were hand-counted in this study. Scorers judged the intent of the voter, counting marks that may not have been counted if the ballots were run through a machine. Because of this, the error rates reported here may represent a best-case scenario; even if participants did not mark ballots as instructed, their marks were counted if their intent could be inferred. This system is similar to manual recounts that are occasionally mandated for auditing purposes or in extremely close races.

Results

Both age and education were treated as continuous variables in these analyses.

Ballot Completion Time

For all four of the voting methods, ballot completion time was measured beginning when the participant entered the voting room and ending when they exited the room.

The overall average ballot completion time for Experiment 1A was 377 seconds (*SD* = 257). Completion times by paired voting methods can be seen in Figure 2. Each pair represents scores from one group of participants. For example, in the two columns labeled "Bubble," the white column represents the ballot completion times on the bubble ballot, and the gray column represents time on the DRE for those same people. Paired t-tests revealed that the difference between the DRE and the lever machine was statistically reliable, $t(17) = 2.87$, $p = .01$, but the other two methods were not.

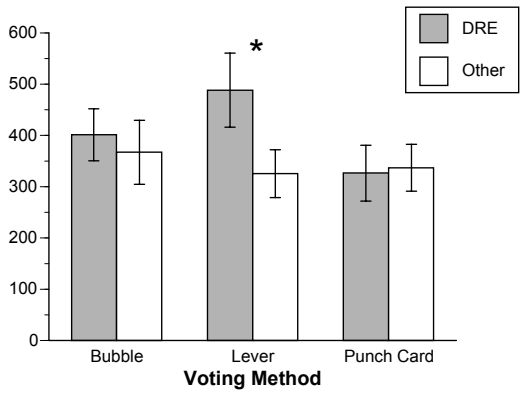

Figure 2. Pairs of mean ballot completion times (in sec) by voting method for Experiment 1A. * $p < .05$.

This effect was not replicated in Experiment 1B; the overall average ballot completion time for 1B was 281 seconds (*SD* = 147) but there were no reliable differences between voting methods.

One other effect that was replicated across both 1A and 1B was that of computer experience. Participants with more self-reported computer experience took less time to vote on the DRE than those with less experience. The correlation between self-reported experience and time was -0.29 (p = .04) for Experiment 1A and -0.28 (p = .03) for 1B. Data from Experiment 1A is presented in Figure 3 (the data from 1B are similar).

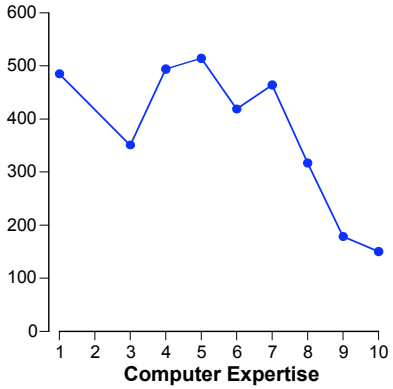

Figure 3. Average DRE ballot completion times for Experiment 1A by computer expertise. 1 = novice, 10 = expert.

In both experiments, there were no reliable effects or interactions involving the information condition variable.

There were two other effects that appeared in only one of the two experiments. Replicating our findings in [5], more educated voters took somewhat less time to vote in Experiment 1A, $F(1, 42) = 4.54$, $p = .04$, but this effect was not statistically reliable in 1B.

Finally, in Experiment 1B, older participants took somewhat longer to vote than younger participants; that is, there was a reliable effect of age on ballot completion times

$F(1, 56) = 13.83$, $p < .001$. This effect was not observed in previous research nor in Experiment 1A.

Errors

In the directed condition, errors were counted when the voter made an incorrect selection (i.e., did not vote as their "slate" instructed). In the undirected condition, errors were counted when a voter's selection was inconsistent with their other vote cast and their intent as given in the exit interview. (For these purposes, the exit interview counts as a third vote). For example, an error in the undirected condition would be if the participant voted for the Republican candidate twice and the Democratic candidate once. The Democratic vote would be counted as an error in this case. Bubble ballots were scored by hand and participants were given credit for marks for which their intention could be inferred, even if the mark may not be counted by machine (e.g. when a participant circled the candidate's name instead of filling in the bubble next to the name).

There are two ways to consider errors: by race and by ballot. On every ballot, there were 27 races and each of these represented a potential for error. Error rates were computed by simply summing the errors committed, and dividing by the total possibilities for errors. The other way to consider errors is by ballot: each ballot can be error-free (i.e., contains zero errors) or have one or more errors.

The per-race error rate in Experiment 1A was 1.6% (SD = 4%) and 3% (SD = 6%) in Experiment 1B. However, there were no differences as a function of experimental condition, nor were there effects of age, education, computer experience, or previous voting experience.

Overall, 27% of the ballots collected in Experiment 1 (both A and B) contained at least one error. However, there was no evidence that any of the four voting methods was more likely to produce a ballot containing an error.

Finally, DREs allow a particularly insidious type of error, termed a postcompletion error [4]. These errors occur when someone leaves out the final subgoal or step in a procedure, and such errors are highly resistant to mitigation efforts [7]. In voting on a DRE, this error occurs when a voter completes all his/her selections, then walks away from the voting booth having failed to activate the final "cast ballot" button. (This is also termed the "fleeing voter" problem.) While this error is possible with the non-DREs, we had never observed this error until introducing DREs in 1A and so did not systematically record it until 1B.[1] In Experiment 1B, 6% of the participants made this error with the DRE. (Votes for these participants were still counted in the assessment of per-race and per-ballot errors.)

Subjective Usability

Average SUS scores and standard deviations for the four voting methods used in Experiment 1A are presented in Table 1. The DREs were reliably superior to the bubble ballot, $t(15) = 2.24$, $p = .04$, and the lever machine, $t(17) = 2.37$, $p = .03$. The difference between the DRE and punch cards approached significance, $t(14) = 1.96$, $p = .07$.

Table 1. Average SUS scores in Experiment 1A

Method	Average SUS	SD
DRE	86.1	16.6
Bubble Ballot	81.3	22.2
Lever Machine	71.5	14.8
Punch card	69.0	22.2

This was replicated and extended in Experiment 1B; satisfaction levels in for each of the four voting methods can be seen in Figure 4. Again, the DRE reliably outperformed the other voting methods: vs. bubble ballot, $t(16) = 4.14$, $p = .001$; vs. lever machine, $t(25) = 4.39$, $p < .001$; vs. punch card $t(19) = 8.46$, $p < .001$.

Satisfaction levels for each voting method were unrelated to their previous voting experience with the method. Satisfaction levels on the DRE were also unrelated to computer expertise. However, the main effect of age also approached significance ($F(1, 56) = 3.9$, $p = .054$), with older adults giving higher SUS scores for both of the methods on which they voted. There were no other significant main effects or interactions on SUS scores.

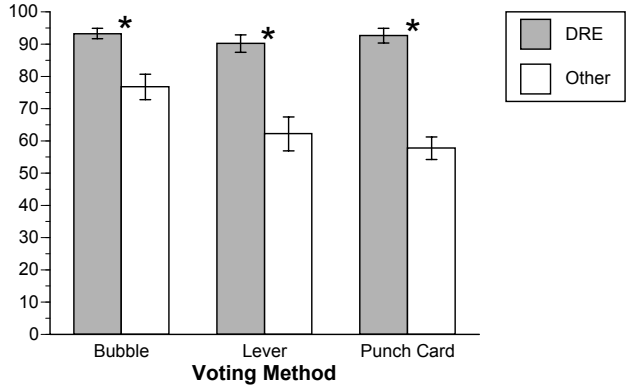

Figure 4. Pairs of mean SUS ratings by voting method for Experiment 1B. * $p < .05$.

Discussion of Experiment 1

When our DRE was compared against older voting methods, the biggest differences were seen in the satisfaction participants felt with each method. The DRE was clearly preferred to the bubble ballot, lever machine,

[1] In fact, the curtain mechanism in the lever machine is so effective at preventing this error that we had to put up a sign reminding participants *not* to pull the "cast" lever so we could record their votes.

and punch card ballot, regardless of any individual characteristics such as age, computer experience, and education. Because of these high satisfaction ratings of the DRE, it is likely that, given experience with the system, voters will not be deterred by the change in technology. These results also suggest that once DREs are adopted, voters may resist any transition back to non-electronic technologies, despite the DREs lack of superiority on objective performance measures.

While voters were most satisfied with their voting experiences on the DRE, this was not due to improvements in error rates or ballot completion times, and the DRE introduced a serious postcompletion error. None of the methods helps solve the problem of voter error, which this and previous studies [5, 10, 12] have shown is a significant concern. In both experiments, more than a quarter of all of ballots contained at least one error, a figure that is much higher than would be desirable in situations like elections with such important outcomes. Simply changing technologies does not necessarily solve problems with user error.

Furthermore, voters with less computer expertise took more time voting on the DRE, though they were not any more or less satisfied with the experience. Voters with more education also took less time voting, an effect previously reported by Byrne et al. [5]. Results such as these can help inform machine allocation decisions. Precincts in which voters are more highly educated or which have more industry requiring computer skills need fewer voting machines since their voters will be able to cast their ballots more quickly. As shown by Mebane [17], in Ohio in 2004, improper allocation of voting equipment can lead to long lines at the polls, which can even affect the outcome of elections.

EXPERIMENT 2

The choice to configure VoteBox to mirror the navigation model used in commercial DREs is useful in that it provides a baseline performance against which alternative interfaces can be compared. In some sense, this also mirrors the presentation in media like punch cards, where the voter must at least page through every intervening race on the ballot to reach a later race.

However, real voters rarely vote in every single available race; the further down the ballot, the less likely it is that a voter will cast a vote in a given race. Why, then, does the typical navigation model require voters to be presented with every race? It is hard to imagine how else one could do this with a paper ballot, but computer interfaces are not so restricted. The primary purpose of Experiment 2 was to compare the "standard" sequential navigation model with a system where users started on an overview screen and could directly jump to individual races, then return to the overview. This is similar to the "hub and spoke" navigation model found on many Web sites.

The obvious question is whether or not this navigation model affects the usability of the system. It is not hard to imagine how this might reduce the time taken to vote, but there may be a cost in terms of votes cast or possibly errors. In Experiment 2, we put this to the test.

Again, this experiment was part of a larger project. Additional information relating to this project is available elsewhere [11].

Methods

Unless stated otherwise, methods for Experiment 2 were identical to those of Experiment 1B.

Participants

Ages of the 64 participants in Experiment 2 ranged from 18-77 years, with a mean age of 50.25 (*SD* = 14.7). 30 were male, and 34 female. The sample was quite diverse in terms of education, ethnicity, and income. Only 8 participants had never voted in a national election before; on average, people had voted in 9.34 national elections. On a Likert scale ranging from 1 to 10 (1 = computer novice, 10 = computer expert), the mean self-rating of computer expertise was 5.7 (*SD* = 2.7).

Design

A new between-subjects variable was added in this experiment—a second VoteBox interface—and a fourth information condition was also added, resulting in a 2 (navigation) x 4 (information condition) x 3 (other voting method used) design. Random assignment was used in all cases.

Half of the participants used a version of VoteBox with a sequential navigation model (used in studies 1A and 1B), while the other half used a direct access version. The sequential DRE interface forced users to page serially through each of the 27 races before casting their vote. Conversely, the direct access interface permitted users to navigate directly to races of their choice, skip undesired races, and move directly to the review screen to cast their ballot at any time.

The three information conditions used in Experiment 1B were undirected, directed with no roll-off, and directed with moderate roll-off. A fourth condition, "directed with additional roll-off," was added in Experiment 2. Motivation for this was in part due to data from a prior study [11] in which roll-off rates in an undirected condition were as high as 78%. The average of the top third of that distribution was .52, which was the number that was used to calculate roll-off rates for the new "directed with additional roll-off" manipulation in Experiment 2.

Materials

Other than the new version of VoteBox, Experiment 2 materials were identical to those from Experiment 1B. The new direct access VoteBox interface was somewhat analogous to a webpage, in that all race titles appeared on

one "home page" or "main menu page" and acted as hyper-links. The main page looked almost identical to the sequential VoteBox review screen.

Procedure

Whereas participants in Experiment 1 voted only twice, Experiment 2 participants voted three times: first with a DRE (either sequential or direct), then with one of the "other" voting methods (paper ballot, punch card, or lever machine), and then again with the same DRE. No exit interviews were conducted in Experiment 2. Since participants now voted three times, errors could be determined by majority rule without the exit interviews that were previously conducted with participants in the undirected information condition.

Results

Ballot Completion Time

Observations more than three interquartile ranges (IQRs) below the 25th or above the 75th percentiles of the distribution were considered outliers and excluded from the analysis; five participants were excluded this way.

The overall average ballot completion time was 290 seconds (*SD* = 183). Mean ballot completion times for each voting method are displayed in Table 4. The difference between sequential and direct DREs was statistically reliable, $F(1, 57) = 4.86$, $p = .03$ (corrected for heteroscedasticity). The mean ballot completion time for the sequential DRE was 442 seconds (*SD* = 397). Participants were markedly faster with the direct DRE, which had a mean ballot completion time of 270 seconds (*SD* = 143). No differences in ballot completion times were found between any other pairs of means.

Table 2. Mean ballot completion times (in seconds) by voting method (with standard deviations) in Experiment 2.

Sequential DRE	Direct DRE	Bubble ballot	Lever machine	Punch card
442.3	269.9	255.7	241.6	239.1
(396.9)	(143.3)	(129.7)	(114.3)	(132.2)

For each of the four information conditions, differences in ballot completion times between voting methods were examined. No timing differences were found within any of the three directed information conditions. The use of a slate (regardless of how many races participants were instructed to skip) did not have differential effects on ballot completion times between any of the voting methods.

However, in the undirected information condition, ballot completion times differed reliably between sequential and direct DREs, $F(1, 13) = 10.47$, $p = .01$. The mean ballot completion time for the sequential DRE was 910 seconds (*SD* = 605). This was much slower than the mean ballot completion time for the direct DRE, which was only 205 seconds (*SD* = 120). This is most likely due to the fact that participants using the direct DRE voted in far fewer contests; as presented in Table 3, those in the sequential DRE condition almost never abstained from a race, but those in the direct condition abstained from nearly half of them. This difference is reliable, $F(1, 14) = 6.91$, $p = .02$, shown in Table 3.

Table 3. Intentional undervote rates by voting method in the undirected condition in Experiment 2.

Sequential DRE	Direct DRE	Bubble ballot	Lever machine	Punch card
.007	.454	.124	.321	.250

Differences between the DRE conditions and the other voting methods were not statistically reliable, nor were the differences between the other voting methods.

Finally, self-reported computer expertise was again negatively correlated with time taken to vote, but only for the direct condition, $r(27) = -0.47$, $p = .014$.

Errors

We differentiated between three types of errors: overvote errors, undervote errors, and wrong choice errors. An overvote error occurs if a voter chooses two candidates for a race in which only a single selection is allowed. This type of overvote error is part of the standard "residual vote" rate and is available in actual elections. The lever machines and DREs (both sequential and direct) prevented participants from making this particular type of overvote error. Thus, only some of our participants even had the opportunity to make this error; none actually did so.

However, a different type of overvote error occurs if a voter makes a selection for a race s/he had originally intended to skip (either due to instructions in the directed information conditions, or personal preference in the undirected condition). The distinction between these two types of overvote errors is an important one; the overvote error rates we report refer to this second definition only and we will refer to these as "extra" votes.

A distinction is also drawn between two types of undervotes: undervote errors and intentional undervotes. An undervote error occurs if a voter fails to choose a candidate for a race in which s/he had intended to vote. An intentional undervote occurs when a voter omits a race on purpose; this is not actually an error. Finally, a wrong choice error occurs when a voter makes a selection other than the one intended.

In cases where the exact nature of an error is irrelevant, the three error types have been combined into a single "any error" error rate. The overall average "any error" error rate was 3.6% (*SD* = 6.7%). When considering each voting method separately, no differences in undervotes, overvotes, or wrong choice errors were found for bubble ballots, lever machines, punch cards, or the sequential DRE. The only

reliable differences between specific error types were seen with the direct DRE. Error rates for each voting method are shown in Table 4. For the direct DRE, the undervote error rate was significantly higher than the extra vote rate, $t(29) = 2.92$, $p = .007$. Undervote error rates were also significantly higher than wrong choice error rates, $t(29) = 2.78$, $p = .01$. Neither previous voting experience nor computer expertise were correlated with any of the various error rates.

Table 4. Error rates by voting method in Experiment 2.

	Extra vote	Undervote	Wrong choice	Any error
Sequential DRE	.000	.002	.012	.013
Direct DRE	.002	.131	.012	.145
Bubble ballot	.000	.002	.002	.004
Lever machine	.000	.006	.011	.017
Punch card	.000	.000	.002	.002

When examining error rates across voting methods, large differences were seen between the sequential and direct DREs. The direct DRE undervote error rate was reliably greater than that of the sequential DRE, $t(29.1) = 2.92$, $p = .007$. Similarly, the direct DRE's "any type" error rate was reliably higher than that of the sequential DRE, $t(29.9) = 2.79$, $p = .009$. (Both tests were corrected for heteroscedasticity.) No reliable differences in error rates were found between paper ballots, punch cards, and lever machines.

As noted in Experiment 1, DREs are additionally plagued by two errors with the unique potential to disenfranchise voters. In addition to the postcompletion error, users of DREs can also cast their vote prematurely by hitting the "cast vote" button before they intended. We did not see this error in Experiment 1, however, in Experiment 2, 8 of 32 people (25%) cast their vote prematurely with the direct DRE, while none did with the sequential DRE. This difference is reliable, $F(1, 62) = 10.33$, $p = .002$. These cases varied in terms of how many choices—if any—had been indicated at the time of accidental casting.

There were no differences in "failed to cast" error rates between the two DREs in this study, though the overall rate of this error was surprisingly high at 9.4%.

As with Experiment 1 and prior work [5, 10, 22], examination of the per-ballot error rate revealed a distressingly high proportion of ballots (32%) contained at least one error. However, no differences between voting methods for number of ballots with at least one error were found.

Subjective Usability

Figure 5 presents the mean SUS rating for each voting method. As in Experiment 1, the sequential DRE was clearly superior to all other methods; vs. bubble ballots, $t(10) = 2.42$, $p = .04$; vs. direct DRE, $t(60) = 4.50$, $p < .001$; vs. punch cards, $t(9) = 3.59$, $p = .01$; vs. lever machines, $t(8) = 3.39$, $p = .01$. However, none of the differences between the other methods were significant.

In general, previous voting experience with a particular technology was not correlated with SUS scores for that technology. The only exception was a significant positive correlation between prior punch card experience and punch card SUS scores, $r(18) = 0.47$, $p = .04$. There were no reliable effects of variables like age and education on SUS scores.

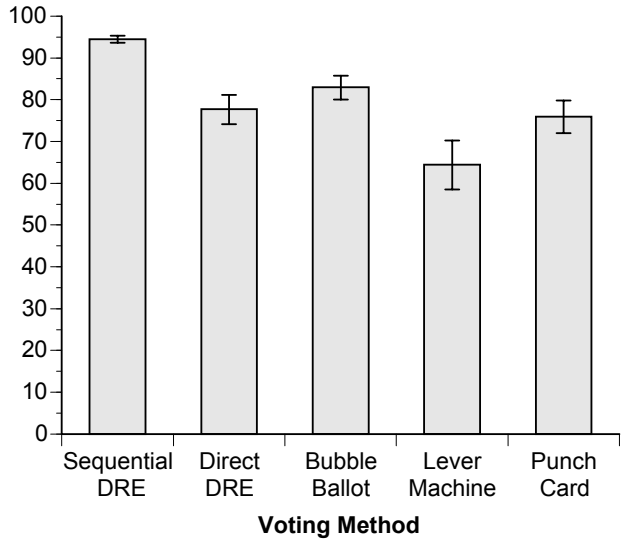

Figure 5. Mean SUS ratings by voting method.

Discussion of Experiment 2

Unlike with Experiment 1, the voting method had a marked impact on objective measures; in particular, the direct DRE was both faster and more error-prone than the sequential DRE. However, even those numbers do not tell the entire story, as the rate of non-erroneous abstention (among participants who had a choice) was also markedly higher for the direct DREs. These differences are certainly large enough that they would likely influence the outcomes of elections if they scaled up to real contests.

Also like Experiment 1, the sequential DRE clearly produced the best subjective usability ratings. This is, in some sense, counterintuitive; UI designers often clearly believe that faster interfaces will be preferred. However, it appears that the inaccuracy of the direct DRE overwhelmed its speed advantage and that became the main driver of subjective scores. Alternately, it may simply be the case that subjective and objective usability measures have little to do with one another.

Although the direct DRE was significantly faster than the sequential, its higher error rate and lower subjective usability ratings are cause for substantial concern. Is it more important that people vote quickly, or that they correctly indicate their choices? How important is voter satisfaction? Presumably one would argue that effectiveness is more

important than efficiency for a voting system, while voter satisfaction may be a tertiary consideration.

These results present another cautionary note; at least in the voting domain, be wary of changes enabled by more flexible technologies. Simply because we now have the ability to support non-paper-like navigation models, and such alternate designs may indeed save users time, it does not mean that such changes are necessarily improvements.

GENERAL DISCUSSION

The goal of these studies was to address the usability of DREs compared to older, more traditional voting methods and to begin to explore the space of what is possible with DREs. It has clearly been assumed by policy makers and others that DREs would be better than the paper ballots, lever machines, and punch cards they have been replacing. Our results indicate that performance on DREs in terms of efficiency and effectiveness is not better than with more traditional methods, and due to the high rate of postcompletion errors it may actually be notably worse. This finding replicated across all experiments and strongly shows that DREs do not necessarily lead to better voting performance as had been assumed. This was a robust effect and although there were some effects of age and education on ballot completion times, in general only computer expertise affected voters' levels of efficiency on the DRE.

No differences in error rates were seen between the "standard" DRE and the older voting methods, but the high frequency of ballots containing errors is cause for concern. The high per-ballot error rates seen here are consistent with previous work on older voting methods [5, 10, 12]. Because the outcomes of elections have such important and widespread impacts, this is clearly problematic.

Although there were no differences in ballot completion times or error rates between the DRE and the other three methods, participants were most satisfied with their voting experience on the electronic voting system. This was indicated by an average SUS score over 90 for this measure in the second experiment. According to suggestions made by Bangor, Kortum, and Miller [1] from their work evaluating a range of technologies, scores above 90 indicate "truly superior products." Using this scale to interpret the SUS scores for the other voting methods shows that the bubble ballot has acceptable SUS scores, while the lever machine and punch card are marginal.

Even though participants with less computer expertise voted more slowly on the DRE, their satisfaction with it was not any less than those users with more computer expertise. The performance on and preference for DREs was not generally affected by age or education levels, though there is some evidence of a digital divide with respect to efficiency. This may be surprising, as some studies (e.g., Roseman and Stephenson [19]) have found a decrease in voter turnout among older people when DREs were used, might have predicted a stronger digital divide effect. We suspect that these effects may be driven by people's expectations about DREs, whereas our study measured only people who had actually experienced the DRE. It is also possible that VoteBox provides a particularly good subjective experience relative to other DREs despite the lack of objective advantages.

It is interesting that voters strongly prefer using the DREs even though their performance is not any better on them. This type of preference versus performance disassociation is not an uncommon finding, but has important implications in elections. Because of the controversy currently surrounding the use of DREs in elections, some groups are calling for a ban on DREs and at least one state (Florida) has reverted from DREs to paper ballots. Although participants do not like paper ballots as much as DREs, they are still generally more satisfied with paper ballots than with lever machines and punch cards. As previous research has reported [5], old-fashioned paper ballots actually work quite well. Most voters can perform at reasonable efficiency and effectiveness levels with paper ballots and are satisfied with the experience. However, paper ballots are seen as inaccessible for many groups of people. For example, paper ballots do not obviously provide voters with disabilities such as low vision or manual dexterity impairments with the ability to vote an independent secret ballot. However, there are, in fact, approaches which can make paper ballots more accessible (e.g. Braille ballots, technologies such as Vote-PAD, http://www.vote-pad.us/), and there are reasons to doubt that current commercial DREs actually do a particularly good job with accessibility concerns [20].

Overall, the usability findings from the current studies show that the use of DREs does not lead to more efficient or effective voting performance, although voters are very satisfied with these electronic systems. The high satisfaction participants feel with the DREs means that citizens may be unhappy about abandoning the new voting systems in the face of security concerns. Even though their performance on the DREs is not any better, voters may fight to keep the systems that they so strongly prefer.

Limitations and Future Directions

The primary limitation of this research is that the DRE used in these studies, VoteBox, is a prototype of an electronic voting system and may not be representative of all DREs. Some systems may be better or worse than the VoteBox system and may return different usability measurements. However, there is little room for improvement in terms of subjective usability, and field studies with commercial DREs are not encouraging in terms of the error rates achieved. Efficiency of commercial DREs is as yet unknown but so far across multiple studies (the ones here and [5, 10, 12]) we have not seen any evidence for strong differences between voting methods on efficiency.

Future studies could include different types of DREs or add features such as touchscreen input or multiple language and font size support to the existing VoteBox system. Another

critical topic is accessibility, both for those who do not read English and for whom physical or perceptual limitations are an issue.

ACKNOWLEDGMENTS

We would like to thank Phil Kortum, Margaret Beier, Amanda Flato, and Michael O'Conner 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 expressed 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

1. Bangor, A., Kortum, P. T., & Miller, J. T. (in press). An empirical evaluation of the System Usability Scale (SUS). To appear in *International Journal of Human-Computer Interaction*.
2. Bederson, B. B., Lee, B., Sherman, R. M., Herrnson, P. S., & Niemi, R. G. (2003). Electronic voting system usability issues. In *Human Factors in Computing Systems: Proceedings of CHI 2003*, (pp. 145–152). New York: ACM.
3. 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* (pp. 189-194). London: Taylor and Francis.
4. Byrne, M. D., & Bovair, S. (1997). A working memory model of a common procedural error. *Cognitive Science, 21*, 31-61.
5. Byrne, M. D., Greene, K. K., & Everett, S. P. (2007). Usability of voting systems: Baseline data for paper, punch cards, and lever machines. In *Human Factors in Computing Systems: Proceedings of CHI 2007,* (pp. 171–180). New York: ACM.
6. Conrad, F. G., Lewis, B., Peytcheva, E., Traugott, M., Hanmer, M. J., & Herrnson, P. S. (2006). *The usability of electronic voting systems: Results from a laboratory study*. Paper presented at the Midwest Political Science Association, Chicago, IL. April 2006.
7. Chung, P. H., & Byrne, M. D. (in press). Cue effectiveness in mitigating postcompletion errors in a routine procedural task. To appear in *International Journal of Human-Computer Studies*.
8. Election Data Services Inc. (2006). *2006 voting equipment study*. Retrieved October 29, 2006 from http://www.eac.gov/voting%20systems/voting-system-certification/2005-vvsg
9. Everett, S. P. (2007). The Usability of Electronic Voting Machines and How Votes Can Be Changed Without Detection. Doctoral dissertation, Rice University, Houston, TX.
10. Everett, S. P., Byrne, M. D., & Greene, K. K. (2006). Measuring the usability of paper ballots: Efficiency, effectiveness, and satisfaction. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.
11. Greene, K. K. (2007). Usability of Electronic Voting Interfaces: Sequential Versus Direct Access. Masters thesis, Rice University, Houston, TX.
12. Greene, K. K., Byrne, M. D., & Everett, S. P. (2006). A comparison of usability between voting methods. In *Proceedings of the 2006 USENIX/ACCURATE Electronic Voting Technology Workshop*. Vancouver, BC.
13. Herrnson, P. S., Niemi, R. G., Hanmer, M. J., Bederson, B. B., Conrad, F. G., & Traugott, M. (2006). *Voters' abilities to cast their votes as intended*. Paper presented at the Workshop on the Usability and Security of Electronic Voting System.
14. 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.*
15. 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.
16. Mebane, Jr., W. R. (2004). The wrong man is president! Overvotes in the 2000 presidential election in Florida. *Perspectives on Politics, 2*(3), 525-535.
17. Mebane, Jr., W. R. (2006). *Voting machine allocation in Franklin County, Ohio, 2004: Response to U.S. Department of Justice letter of June 29, 2005*. Retrieved February 18, 2007 from macht.arts.cornell.edu/wrm1/franklin2.pdf
18. Nichols, S. M., & Strizek, G. A. (1995). Electronic voting machines and ballot roll-off. *American Politics Quarterly, 23*(3), 300-318.
19. Roseman, G. H., Jr., & Stephenson, E. F. (2005). The effect of voting technology on voter turnout: Do computers scare the elderly? *Public Choice, 123*, 39-47.
20. Runyan, N. (2007). *Improving access to voting: A report on the technology for accessible voting systems*. Retrieved March 1, 2007 from http://demos.org/pubs/improving_access.pdf
21. Sandler, D., & Wallach, D. S. (2007). Casting votes in the Auditorium. In *Proceedings of the 2nd USENIX/ACCURATE Electronic Voting Technology Workshop*. Boston, MA.
22. Traugott, M. W., Hanmer, M. J., Park, W.-H., Herrnson, P. S., Niemi, R. G., Bederson, B. B., & Conrad, F. G. (2005). *The impact of voting systems on residual votes, incomplete ballots, and other measures of voting behavior*. Paper presented at the annual conference of the Midwest Political Science Association, Chicago, IL, April 7-10.
23. Wand, J. N., Shotts, K. W., Sekhon, J. S., Mebane, W. R., Jr., Herron, M. C., & Brady, J. E. (2001). The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida. *American Political Science Review, 95*(4).
24. Yee, K. P., Wagner, D., Hearst, M., & Bellovin, S. M. (2006). Prerendered user interfaces for higher-assurance electronic voting. In *Proceedings of the USENIX/ACCURATE Electronic Voting Technology Workshop*, Vancouver, B.C.