USENIX Journal of Election Technology and Systems (JETS)
Volume 1, Number 1 • August 2013

How To Build an Undervoting Machine: Lessons from an Alternative Ballot Design

KRISTEN K. GREENE, RICE UNIVERSITY*
MICHAEL D. BYRNE, RICE UNIVERSITY
STEPHEN N. GOGGIN, UNIVERSITY OF CALIFORNIA

Despite the importance of usability in ensuring election integrity, it remains an under-studied aspect of voting systems. Voting computers (a.k.a. DREs) offer the opportunity to present ballots to voters in novel ways, yet this space has not been systematically explored. We constructed a DRE that, unlike most commercial DREs, does not require voters to view every race, but instead starts at the "review screen" and lets voters directly navigate to races. This was compared with a more traditional, sequentially-navigated, DRE. The direct access navigation model had two effects, both of which were quite large. First, voters made omission (undervote) errors markedly more often. Second, voters who were free to choose who to vote for chose to vote in substantially fewer races. We also examined the relationship between the true error rate—which is not observable in real elections—and the residual vote rate, a measure of effectiveness commonly used for real elections. Replicating the findings of [Campbell and Byrne 2009a], the mean residual vote rate was close to the mean true error rate, but the correlation between these measures was low, suggesting a loose coupling between these two measures.

1. INTRODUCTION

One of the most obviously desirable qualities of a voting system is that it accurately measures the intent of the voters. With the advent of direct recording electronic (DRE) voting machines, the concern for accuracy has been raised in the form of discussions about computer security, since most commercial DREs are susceptible to various form of attack, including computer viruses, e.g., [Ohio Secretary of State 2007]. While security concerns have led many states to move away from DREs towards optical scan paper-ballots, there are several compelling reasons why DREs remain an important research area. For example, with electronic technology, there exists significant potential for accessibility accommodations that traditional paper-based systems cannot offer; there will always be a need for systems that support advances in accessible technology. From a much broader research perspective, DREs offer the opportunity to present ballots to voters in novel ways, yet this space has not been systematically explored. Without understanding how specific design features—such as navigation style—impact usability of known voting systems, we will have no baseline data against which to evaluate emerging and future voting methods. For obvious reasons, experimental manipulations of potentially significant ballot design features are neither feasible nor desirable in real-world elections, lest the will of some voters be more accurately reflected than others. By conducting a mock election in a controlled setting, we were able to gather both objective and subjective usability data across multiple voting systems (two DREs with different navigation styles, paper ballots, punch cards, and lever machines) at a much finer level of granularity than is possible in a real election. The laboratory environment also allowed a comparison of true voter error with the indirect measure commonly used in political science, i.e., the residual vote.

Studies outside the laboratory, such as archival studies of voting records from enormous samples of counties across multiple elections utilize "residual votes" as the measure for error. [Stewart 2006] regards it as "a measure of voting machine accuracy that was first championed by the Caltech/MIT Voting Technology Project in 2001 and has been regularly used ever since..." These residual vote rate studies, while important, often neglect to examine the different types of voter errors, and utilize different methodologies for counting errors.

For instance, [Ansolabehere and Stewart 2005] found that the types of voting technologies in use have an impact on under- or overvote occurrences, or the residual vote rate. Ansolabehere and Stewart found clear differences between technologies in terms of the residual vote rate, with the results differing based on the location of the race on the ballot. Specifically, they found that punch

This research was supported by the National Science Foundation under grant #CNS-0524211 (the ACCURATE center).
*Now at the National Institute of Standards and Technology (NIST).

cards created the most residual votes for presidential races, yet lever machines produced the highest rates for down-ballot races. Optical scan ballots perform best at presidential, or top-of-the-ballot races, while there is a three-way tie between optical scan, DRE, and traditional paper ballots for lowest residual vote rate among down-ballot races. Importantly, Ansolabehere and Stewart (p. 378), unlike most other studies in Table 1, note the importance of a race's place on the ballot, stating that "which machine is 'best' appears to depend, in part, on which race you are studying." While [Ansolabehere and Stewart 2005], as well as [Traugott et al. 2005] note the differences in residual vote rate among different races on the ballot, neither study focuses on these differences.

These results highlight an important problem with using the measure of residual votes for the metric of effectiveness in studying voting machine usability. Because residual vote rates include both undervotes and invalid overvotes (or spoiled ballots), they are sensitive to the effects of ballot roll-off, and do not provide as precise a measure as possible in the laboratory. While ballot roll-off rates, estimated for DREs and voters in such articles as [Nichols and Strizek 1995] or [Kimball et al. 2001] could be controlled for in the residual vote rate, the residual vote rate is often not completely separated into its constituent parts. Any "switched," or "wrong choice" ballots, in which a vote was cast for any candidate(s) other than the candidate(s) the voter intended, are not captured as error at all. As [Kimball and Kropf 2008] noted, residual vote rates on propositional measures are much higher than races at the top of the ballot, and while this may be driven by rational undervoting, aggregate data does not allow the separation of the residual vote into intentional and unintentional undervotes, let alone wrong choice votes and overvotes.

Furthermore, in massive archival studies of voting records using residual vote rates, reporting methods differ based on county and state, and often different rules must be applied to calculate the residual vote rate based on the information given by the election officials. In nearly all studies, the measure of an undervote, is merely whether a voter left the first race on the ballot blank. Often, this is the Presidential race in a Presidential election year. While it is likely that many who leave this race blank likely do so unintentionally, it is still feasible voters abstain intentionally, which would then be measured as error.

Even if the residual vote rate is an accurate representation of the errors produced by voters in the election process, traditional studies of election technology in political science typically neglect the other two metrics of usability, efficiency and satisfaction. Usability is a multifaceted problem, as described clearly in a 2005 report from the National Institute of Standards and Technology (NIST) [Laskowski et al. 2005], which recommends voting systems be evaluated by the ISO standard metrics of usability: effectiveness (for voting systems, accuracy), efficiency, and satisfaction. While some studies, such as [Stein et al. 2008] have examined the time it takes a voter to cast a ballot in the field, no comprehensive field studies have been able to record measures for all three metrics for the same voters. While this difficulty is due to election law and the privacy of the voter, alternative methodologies are available for studying voting machine usability, such as in the laboratory.

In 2006, Byrne, Everett, Greene and colleagues began collecting baseline usability data for traditional voting methods. Across the series of three usability studies [Everett et al. 2006; Greene et al. 2006; Byrne et al. 2007], they gathered baseline usability data for multiple forms of paper ballots, punch cards, and lever voting machines. Their research showed that on all technologies, error rates were high, across these three studies and multiple voting methods, between 11 and 26% of the ballots contained at least one error and error rates per-race were generally between 1 and 4%. Furthermore, the differences between these technologies on objective measures of usability are not large, though they found some evidence that punch cards and lever machines are more error-prone for some populations, e.g., older voters tend to have more trouble with lever machines [Byrne et al. 2007]. However, these studies showed a small but consistent advantage for paper ballots in terms of subjective usability.

In 2008, Everett and colleagues [Everett et al. 2008] report on laboratory studies that compared the usability of a new prototype DRE versus traditional voting methods (paper ballots, punch cards, and lever machines). While there were few differences between the DRE and the older voting methods on efficiency or effectiveness, participants were substantially more satisfied when using the DRE. The bubble ballot received high SUS scores as it had in prior research, but

with the addition of a DRE, the bubble ballot was no longer participants' favorite voting method. Despite the large subjective advantage for DREs, participants' objective performance was not any better with the DRE. This disassociation between subjective and objective usability has been seen in a number of domains [Bailey 1995, Neilsen 1995], and may have ramifications for election officials who decide to return to traditional methods after having previously adopted DREs.

[Herrnson et al. 2008] reports another substantial usability evaluation; they evaluated six electronic voting systems and four verification systems using expert reviews, a laboratory experiment, and a field study. The six commercial DREs included an ES&S Model 100, a Diebold Accuvote-TS, an Avante Vote-trakker, a Hart InterCivic eSlate, a Nedap Liberty Vote, and a Zoomable prototype system developed specifically for this research. Participants were asked to read a voter guide and circle their choices, then use the marked-up guide while voting. However, even though they were asked to make their own choices, for some races they were instructed on who to vote for, and were also asked to intentionally omit one down-ballot race, change a vote, and do a write-in.

When using a standard office-bloc ballot design and performing no special tasks, voters cast their ballots accurately over 97% of the time. In this scenario, there was little difference between voting systems in terms of accuracy. However, attempting to select multiple candidates, change a selection, and vote straight-party ticket caused accuracy to drop sharply, down to the range of 80-90%. Most importantly, differences in accuracy between voting systems were then seen. For example, accuracy for the Hart InterCivic system was only 72% and accuracy for the Avante was much worse, at 50%. Soberingly, roughly 20% of voters cast ballots that were not completely accurate. Despite this, all six DREs were viewed favorably by voters, who had relatively high confidence with the commercial DREs. In fact, voters were more confident with touch screens than with paper, and also judged touch screens as more trustworthy than paper. [Herrnson et al. 2008]

Unfortunately, the 2008 study by Herrnson et al. did not evaluate the efficiency of the various systems; ballot completion times were not recorded. An additional study limitation was the use of a non-standardized questionnaire to assess voter satisfaction. Use of a standardized instrument, such as the System Usability Scale [Brooke 1996], would have facilitated comparison of results across studies and technologies. Finally, the experimental procedures used were somewhat artificial. Having participants make their own selections, yet instructing them to change those selections in specific ways, is not a task scenario that is representative of what most voters experience in a real election. Despite these methodological limitations, the 2008 study by Herrnson et al. is an important landmark it the study of voting system usability and provides an important reference point for future research, including its focus on the separation of types of voter errors to include undervotes, overvotes, and "wrong choice" votes.

So, while there has been some recent research which has advanced our understanding of voting system usability, a great deal more research is needed. Our current research is aimed at two issues: first, we want to widen the space of inquiry into DRE design. The [Herrnson et al. 2008] research demonstrates that the differences between various DREs, while meaningful, are mostly not large (with some exceptions), and some of these differences are almost certainly exacerbated by the requirement that voters make a selection and then change it. While looking at commercial DREs is certainly valuable, it is also limited because most commercial DREs use similar designs for many functions. For instance, most DREs use touchscreen-activated textual buttons rather than alternatives like handwriting recognition, presentation of images rather than text, or other non-traditional designs. These design features are likely motivated by a desire (sometimes driven by state laws regarding ballot design) to keep the voting experience isomorphic to more traditional presentation on a paper ballot, but there are aspects of the paper ballot experience that are not captured in commercial DREs. In particular, most major commercial DREs present the ballot sequentially: that is, the race for Representative follows the race for Senator which follows the race for President. While a paper ballot typically also presents races in this order, voters can typically see all (or most) of the races simultaneously and can make their selections in whatever

order they wish.[1] Does giving voters the same flexibility with a DRE alter usability? It is clear that navigability is a critical factor in the usability of other systems such as Web sites—perhaps this applies to ballots as well.

Second, we want to further clarify the relationship between residual votes and true error rates through experimental work. Residual vote rate is an indirect measure of error; in an experiment, error can be more directly assessed. How well does residual vote rate compare to true error rate? Some work on this has already been published. e.g., [Campbell and Byrne 2009] showing that while the overall average residual vote and true error rates may be similar, the variability in one is not mirrored by the other. That is, when ballots are aggregated, the total residual vote proportion and the true error proportion may be similar, but this is not true at the individual ballot level; the correlation between the residual vote rate and the true error rate was found to be low. We wanted to further explore this relationship.

2. EXPERIMENT

2.1 Method

Participants. Sixty-four participants (30 male, 34 female) from the greater Houston, TX area were recruited through newspaper and online ads for participation in a mock election study. Ages ranged from 18 to 77 years, with a mean age of 50.3 (SD = 14.8). There was reasonable variability in terms of participant ethnicity, annual income, and education.

Design. The primary independent variable of interest was navigation style for the DRE, which was manipulated between subjects with random assignment. In addition to the sequential navigation style used in most commercial DREs and other studies—e.g. [Everett et al. 2008]—a "direct access" navigation style was also used. Half of participants voted with a sequential DRE; the other half voted with a direct access DRE. Each participant voted twice with either the sequential or the direct access DRE system. Both DREs had the same 27 contests, comprised of 21 candidate races and six propositions. The direct access DRE was similar to a webpage, in that all race titles appeared on its Main Page and acted as hyperlinks; see Figure 1. From the Main Page, a voter could click on the race titles to pick and choose exactly the races s/he wanted to see, with the option to skip races and go straight to the Record Vote screen at any time; see Figures 2 and 3 for a candidate race screen and the Record Vote screen, respectively. Note that merely moving from the Main Page to the Record Vote screen did not cast the ballot. The Record Vote screen presented the voter with buttons allowing the voter to either return to the Main Page or to cast his/her vote by selecting the "Record Vote" button (Figure 3).

In sharp contrast to the "pick and choose" navigation model of the direct access DRE, the sequential DRE forced voters to page sequentially through every race on the ballot—followed by a Review Choices screen much like the direct access DRE's Main Page—before they could get to the Record Vote screen. While the overall navigation schemes differed substantially between DREs, their Record Vote screens were nearly identical.

In addition to voting twice on a DRE, each subject voted on one of three other non-DRE methods: paper ballots, punch cards, or lever machines. Participants voted on the non-DRE system in between the two votes with the DRE. The assignment of participants to each level of this between-subjects variable was random.

A third independent variable was termed "information condition" and had four levels; participants were randomly assigned to one of these four. One level was termed "undirected;" in this condition participants were given a fictional voter guide modeled after the guides provided by the League of Women Voters, identical to the one first used in [Greene et al. 2006]. This guide

[1] It is, in principle, possible to present all of the races simultaneously on a DRE, but doing so for many U.S. elections would require extremely large, and therefore prohibitively expensive, computer displays. Machines such as the AVC Advantage and the Nedap LibertyVote do in fact do this, using a display of lights paired with paper-printed text on a large face, controlled by a computer. These machines, while DREs, largely resemble other styles of traditional voting machines, such as lever-based machines, as the user interface is not a traditional computer display.

provided specific information about each candidate for office, in addition to arguments for and against the propositions on the ballot. Participants read this guide and then chose for themselves the candidates and propositions to vote for. The remaining three conditions were "directed" in that participants were given a list of candidates to vote for. In the "directed with no roll-off" condition, this list was complete, instructing voters to cast a vote in every race on the ballot. Other levels instructed voters to abstain in some races. "Directed with moderate roll-off" was based on the [Nichols and Strizek 1995] data on roll-off rates. We used the average roll-off rates for jurisdictions with electronic voting machines for the various offices: for federal offices (9.75%), for state offices (12.00%) and for county offices (16.37%). This was designed to more closely mimic real-world voting patterns in which people do not vote for every race on the ballot. The "directed with additional roll-off" condition used lists where each race had a 52% chance of being omitted in order to simulate voters who only vote in a few races in each election.

A more complete listing of the materials including screen shots for most of the DRE screens, the full ballot, the full voter's guide, and example slates can be found at <http://chil.rice.edu/research/jets13/>.

The three dependent variables measured were ballot completion time, errors, and satisfaction. Ballot completion time was measured in seconds, and indicated how long it took participants to complete each of their three ballots. In the directed information conditions, errors were recorded if a participant chose a candidate other than the one they were instructed to select, if they failed to make a choice for a race in which they were supposed to vote, or if they made a choice in a race they were supposed to skip. In the undirected information condition, errors were recorded if there was a discrepancy between responses on a subject's three measures. Finally, satisfaction was measured using the System Usability Scale (SUS) [Brooke 2006], a ten-item battery of questions regarding subjective usability. Participants filled out three separate SUS questionnaires, one for each ballot they completed.

Errors were further broken down by error type: an error was a *wrong choice error* if the option selected was the incorrect one (e.g., a vote for Alice when a vote for Bob was intended). If a voter chose a candidate for a race s/he had planned to (or was instructed to) omit, this was considered an *extra vote error*. If a voter omitted a race that required a vote, this was considered an *omission error*. Finally, on the paper or punchcard systems, voters could vote for more choices than allowed, this was an *overvote error*.

In contrast with an undervote error, an *intentional abstention* occurred when a voter correctly abstained from voting in a race s/he had intended to skip. This is not an error, but is an important measure of voter behavior, as these would be counted as errors in the residual vote rate.

Procedures. Participants were first given a set of written instructions explaining the purpose of the study, accompanied by either a slate of candidates they were directed to vote for or a fictional voter guide. If given the guide, the participants were given as much time as they wanted to familiarize themselves with the political candidates and propositions. Once the participant indicated he or she was ready to begin, several key instructions were reiterated verbally. Participants were instructed to vote for the same candidates on all ballots.

Participants voted first on a DRE, then on one of the three non-DREs, then again on the DRE. This did away with the need for an exit interview in the undirected information condition. Finally, the participants filled out a longer questionnaire about their voting history and demographic information.

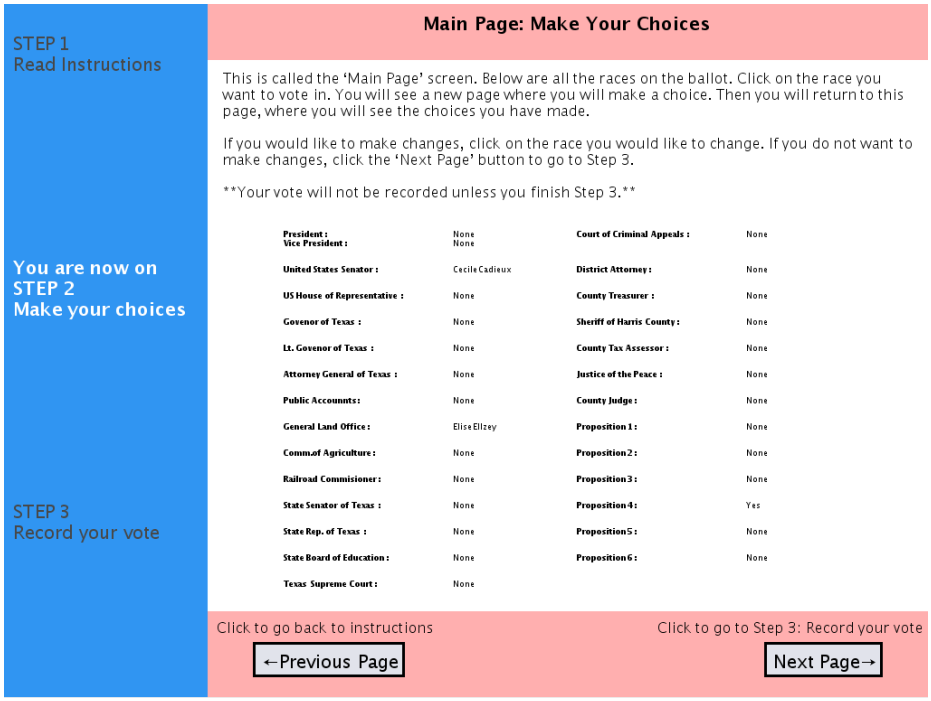

Fig. 1. "Main Page" for the Direct Access DRE.

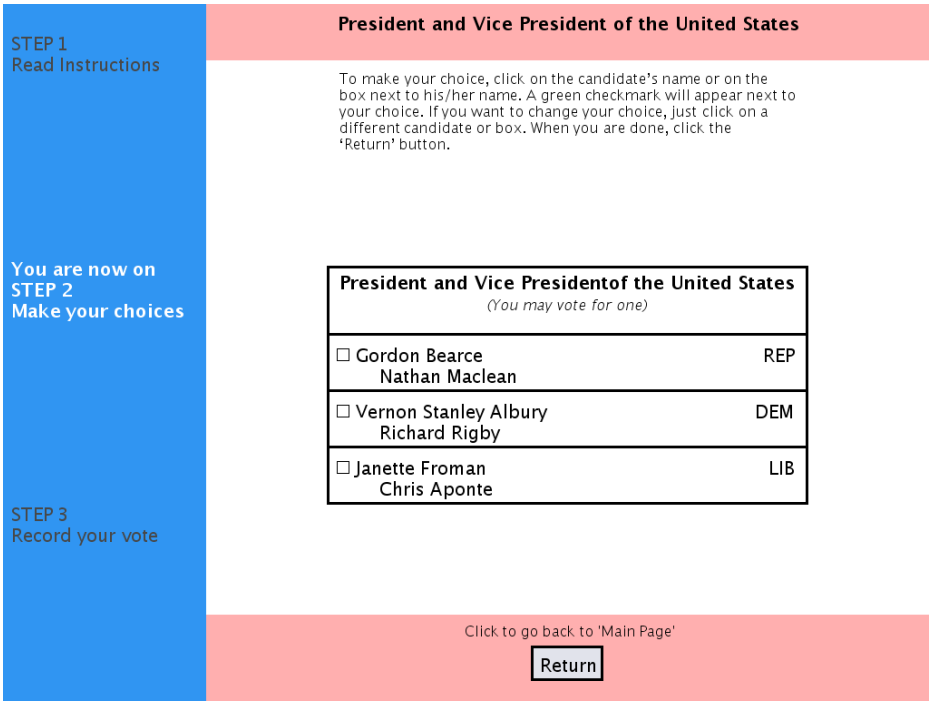

Fig. 2. Candidate Race Screen for the Direct Access DRE.

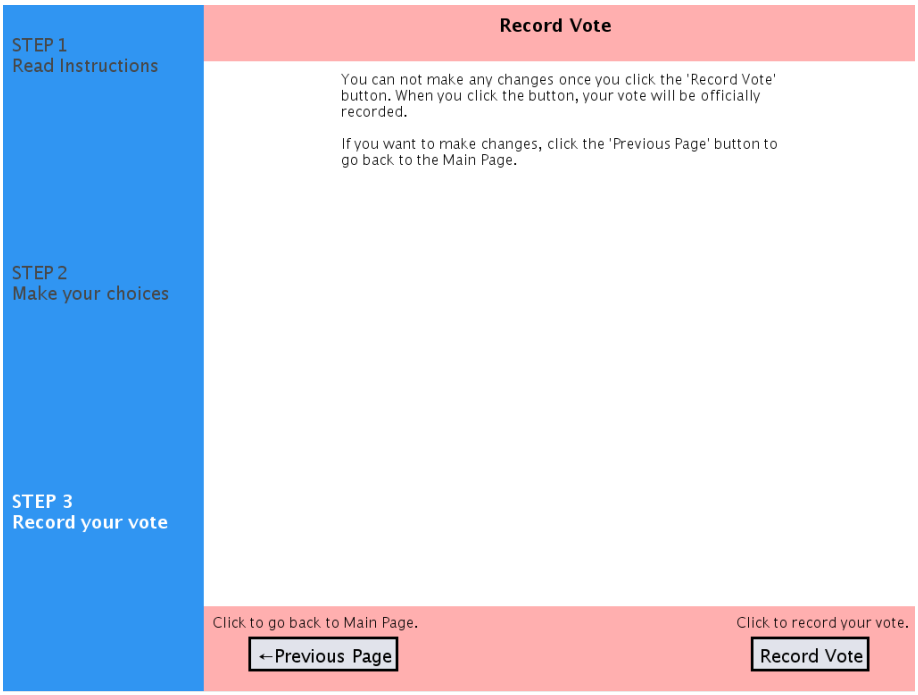

Fig. 3. Record Vote Screen for the Direct Access DRE.

2.2 Results

Efficiency. Data from six participants were not included in the efficiency analyses due to being outliers, defined as observations falling outside three interquartile ranges (IQRs) from the 25th or 75th percentiles of the distribution of ballot completion times. The most striking effects of navigation type on efficiency were seen in the undirected information condition, where ballot completion times for the sequential DRE were over twice as long as times for the direct access DRE: 453 seconds (SD=123) versus only 205 seconds (SD=119) . The interaction between navigation type and information condition, depicted in Figure 4, was statistically reliable, $F(3, 34) = 5.26$, $p = .004$, as was the main effect of navigation type, $F(1, 34) = 11.4$, $p = .002$. This result is inconsistent with previous results in that effect of voting system have generally been absent (or small); however, this result should not be considered in isolation from results on efficiency and intentional undervotes.

While navigation type had a significant impact on efficiency, type of non-DRE voting method used did not.

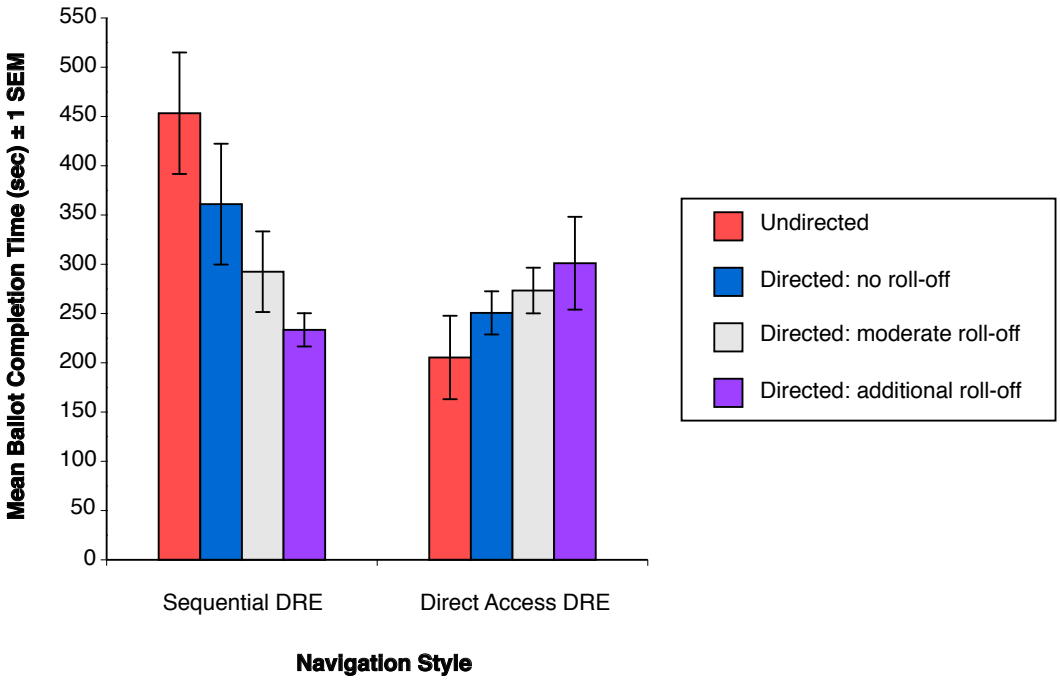

Fig. 4. Effects of DRE navigation style and information condition on ballot completion times

Effectiveness. Data from six participants were not included in the effectiveness analyses due to being outliers, defined as having greater than 15% errors on all three ballots they completed. This definition is in keeping with previous work [Byrne et al. 2007]. No overvotes were observed. Error rates for each voting technology are presented in Table 1.

Table 1. Mean Error Rates by Voting Technology and Type of Error

Technology	Extra vote	Omission	Wrong choice	Total error
Sequential DRE	0.0%	0.2%	1.1%	1.3%
Direct Access DRE	0.2%	13.1%	1.2%	14.5%
Bubble ballot	0.0%	0.2%	0.2%	0.4%
Lever machine	0.0%	0.6%	1.1%	1.7%
Punch card	0.0%	0.0%	0.2%	0.2%

Most interesting was the extremely high undervote error rate for the direct access DRE, which was about 13%, as opposed to only 0.2% for the sequential DRE. The main effect of navigation type on error rates was reliable, $F(1, 17) = 7.39$, $p = .02$. The abnormally high undervote error rate for the direct access DRE can be explained by the number of people who cast their ballots prematurely with the direct access DRE. In comparison with more traditional voting methods, DREs offer opportunities for voters to commit two particularly severe errors. A voter can fail to cast their vote entirely by not pressing the "Record Vote" button at all. This happens in real elections, and has been termed the "fleeing voter" problem [Felten 2009]. In a real election, when a voter fails to cast their vote, there is still a chance that the next voter or a poll worker will cast it for them. Of course, this is a function of the behavior of the next voter, the poll worker, and voting laws in the jurisdiction. A malicious next voter could see that the machine was left in an un-cast

state, change all the selections, and effectively vote twice, though this is illegal in most jurisdictions. Alternatively, the voter could get a poll worker, who in some jurisdictions is allowed to cast the vote, and in others is required to cancel the vote, therefore disenfranchising the voter who fled.

In this study, if a participant failed to cast their ballot, we cast it for them and counted their choices as intended. In contrast with failing to cast a vote at all, a voter can cast their vote prematurely, by pressing the "Record Vote" button too soon. This is also quite problematic, for when a vote is cast prematurely, voters irreversibly rob themselves of the opportunity to vote in some or all of the races on a ballot.

Of the 32 people who used the sequential DRE, only two people failed to cast their vote (6.3%), and not a single person cast their vote too soon. However, of the 32 people who used the direct access DRE, four people failed to cast their vote (12.5%), and eight people cast their vote too soon (25%). The number of races erroneously omitted due to premature ballot casting varied, and in several cases, no choices at all had been made when the ballot was cast. The direct access DRE had a significantly greater "cast too soon" error rate than did the sequential, $F(1, 62) = 4.33$, $p = .002$.

Intentional Abstentions. In a real election, it is impossible to discern whether an omission was an error on the part of the voter, or whether it was intentional. The controlled nature of this mock election allowed a distinction to be made between undervote errors and intentional undervotes. The intentional abstention rates reported here are for the undirected information condition only. Such rates are not of interest for the directed information conditions, in which participants were not allowed to make their own decisions regarding abstentions.

There was quite a large disparity in intentional undervote rates between the sequential and direct access DREs. Participants who used the sequential DRE almost never abstained from a race, resulting in an intentional abstention rate of 0.7% for those voters. In sharp contrast, those who used the direct access DRE abstained from nearly half of all races, with an intentional undervote rate of 45.4%, a dramatic difference, $F(1, 14) = 6.94$, $p = .02$.

Residual Vote versus True Error Rate. Of great interest was the relationship between the study's true error rate (a direct measure of effectiveness) with what would have been reported as the residual vote (an indirect measure) in a real-world election. The residual vote in our analysis is comprised of overvote errors, omission errors, and intentional abstentions. The residual vote does not include any information about wrong choice errors because it is impossible to identify such errors without knowing voter intent, something which is impractical to do in real elections due to privacy concerns. For the same reason, the residual vote cannot differentiate between omission errors and intentional abstentions. The residual vote rate was then compared to our measure of the true overall error rate, i.e. the "total" error rate, which was comprised of extra vote errors, overvotes, omission errors, and wrong choice errors combined. Comparisons between the residual vote rate versus the true error rate were done for the undirected information condition only, as voters the directed conditions did not have the option to choose their own abstention opportunities. It is common to compare accuracy between voting systems by focusing on the top-ballot residual vote rate, since intentional abstention rates are known to increase for down-ticket races. Therefore, comparisons between what would have been reported as the residual vote rate versus the true error rate are shown separately for the Presidential race (Figure 5) and all other down-ballot races (Figure 6). Given the extremely large intentional abstention rate for those voters using the direct access DRE, residual vote rates and error rates are also broken down by navigation type.

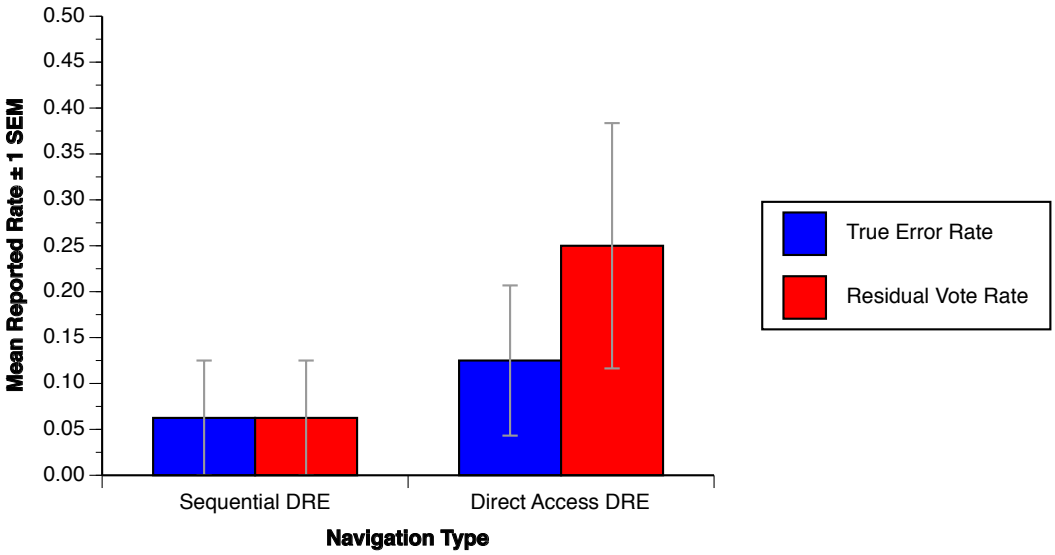

Fig. 5. Top-ballot (e.g., Presidential) residual votes and true error rates for the sequential navigation DRE and the direct access navigation DRE

The correlation between the residual vote and true error rate for the Presidential race was not significant for sequential navigation, $r(6) = -.14$, $p = .74$, nor was the difference between means reliable, $t(7) = 0.00$, $p = 1.00$. A similar pattern of results for the presidential race was observed for direct access navigation: the correlation between the residual vote and true error rate was not significant, $r(6) = .41$, $p = .32$, nor was the difference in means, $t(7) = 1.00$, $p = .35$.

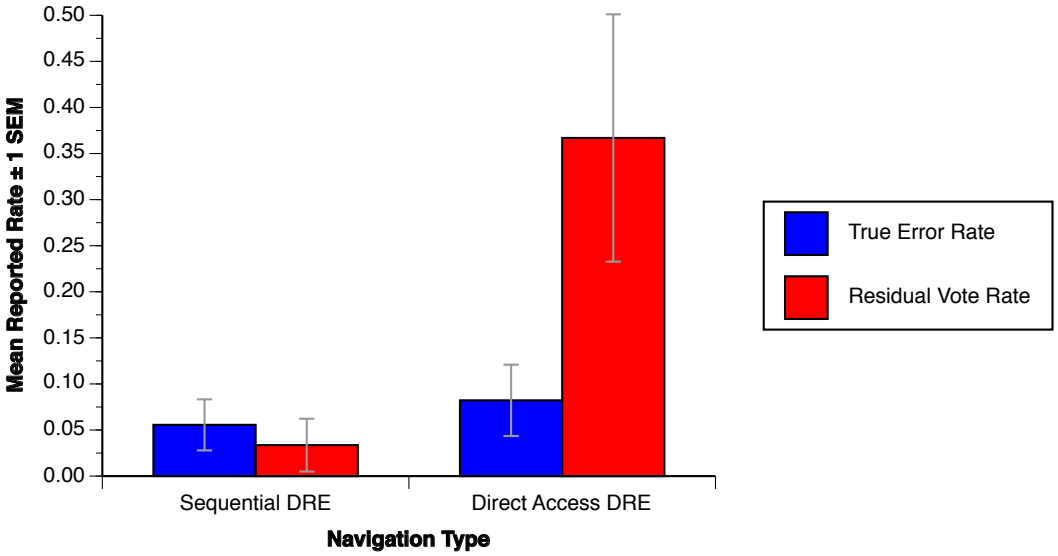

Fig. 6. Down-ballot residual votes and true error rates for the sequential navigation DRE and the direct access navigation DRE

Similarly, the correlation between the residual vote and true error rate for the remaining down-ballot races was not significant when navigation type was sequential, $r(6) = .22$, $p = .60$. Again, the difference in means was not reliable, $t(7) = .43$, $p = .68$. When navigation type was direct access, the correlation between the down-ballot residual vote and true error rates was not significant, $r(6) = .16$, $p = .70$, nor was there a significant difference in means, $t(7) = 1.64$, $p = .15$.

These non-significant correlations suggest the residual vote may not be as tightly coupled to the true error rate as has often been assumed.

Satisfaction. Data from two participants were excluded due to being outliers, defined as any point falling outside three IQRs from the 25th or 75th percentiles of the SUS scores distribution.

The sequential DRE consistently received higher SUS scores than did the direct access DRE, see Figure 7; the effect of navigation type on SUS scores was reliable, $F(1, 38) = 9.53$, $p = .004$. Although differences in SUS scores were seen as a result of changing DRE navigation style, information condition did not significantly impact subjective usability.

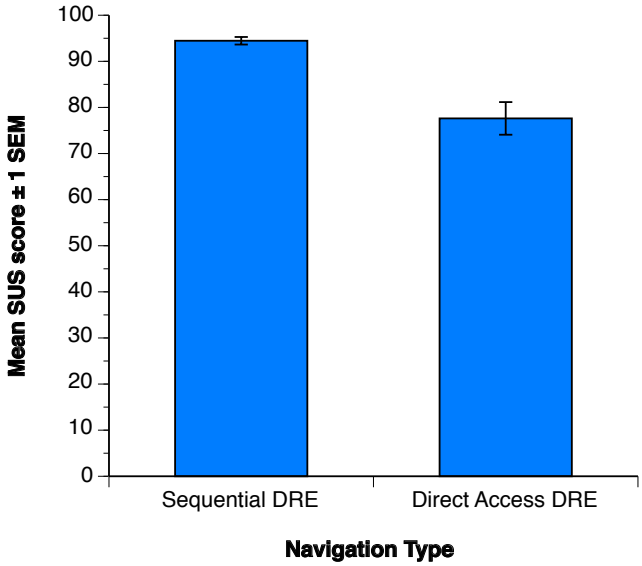

Fig. 7. Mean subjective usability (SUS) scores for sequential vs. direct access DREs

Consistent with multiple other studies, the sequential DRE received the highest mean SUS score of any voting method; ratings for the non-DRE voting methods overall were lower than ratings for the DREs. This is depicted in Figure 8. Unsurprisingly, all three pairwise differences between the DRE and the other methods were statistically reliable, $p < .05$ in each case.

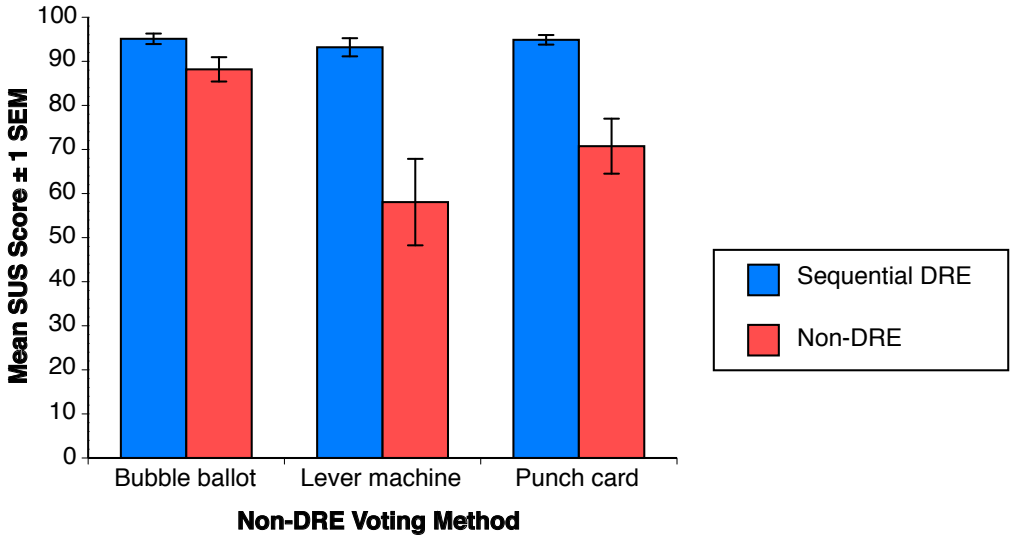

Fig. 8. Mean SUS scores for sequential DRE versus non-DRE voting methods

3. DISCUSSION

Exploration of alternatives in terms of ballot presentation has the potential to be valuable. From a user interface perspective, many DRE systems are essentially electronic implementations of paper ballots and do not take advantage of alternatives made possible by the electronic medium, such as alternate navigation models. We attempted to make a foray into this space, and examined a navigation style more like the Web. However, manipulation of a that single design feature—how the ballot was navigated—resulted in substantial effects on both subjective and objective usability. The lesson here is that changes in one aspect of the voting interface can have far-reaching consequences in terms of overall system usability. This should raise serious concerns about other purportedly small changes in voting procedures—such changes can have substantial impacts and no such changes should be deployed without being thoroughly tested for their effects on usability. For instance, there are a variety of proposals in the literature for voting systems with "end to end" verifiability (e.g., Prêt à Voter [Ryan, et al. 2009]) that introduce changes to the voting procedure. The usability ramifications of these changes have not been thoroughly tested, and it is not unreasonable to wonder if they have similar impacts.

In our case, intentional abstention rates, efficiency, effectiveness, and satisfaction were all greatly impacted simply by changing the DRE navigation style from sequential to direct access. When allowed to choose candidates for themselves in the undirected information condition, participants with the sequential DRE voted in almost every race, whereas participants using the direct access DRE abstained from nearly half of all races. Voting in so few races made the direct access DRE significantly faster than the sequential, but this increase in speed came at a tremendously high cost in accuracy. The direct access DRE had a much greater omission error rate than any of the other voting methods, and was plagued by a new type of error: voters frequently cast their ballots prematurely with the direct access DRE. While some voters cast their ballot too early, others failed to cast their ballot at all, replicating the real-world "fleeing voter" phenomenon in the laboratory.

That the results were so large was certainly unexpected. The change in navigation model generated substantial amounts of undervotes, both erroneous omissions and intentional abstentions. It is possible that some kind of visual cue to highlight races that had not yet been voted, such as the bright orange coloration employed by [Campbell and Byrne 2009a] to improve anomaly detection might also better alert voters to those races that they had not yet voted in. However, this seems somewhat unlikely, as when the voter first starts using the interface, *all* races would be flagged this way, so nothing would stand out. This is most likely to help only when there are a few isolated incomplete races. This does not imply that all visual UI manipulations would in principle be ineffective, but it seems unlikely that there is a simple visual fix on the main screen.

We strongly suspect that it is the sequential presentation that forces voters to see every single race that encourages voters to consider each race, making them much more likely to choose to cast a vote. However, this raises issues of scale. Our ballot was designed to represent an average length ballot in a U.S. election, but there is considerable variance in this measure and in some jurisdictions, ballots can be more than 100 contests long. How such a long ballot would be handled in a direct access DRE is not entirely clear. In order to maintain a readable font size, races would either have to be broken across multiple pages or the interface would have to support scrolling. Recent research [Campbell, 2013] suggests that a paginated display is likely to be better, but neither of these circumstances would be ideal.

One of the methods deployed in some jurisdictions to help alleviate the time taken for long ballots is the straight-party voting (SPV) option. This is currently available in roughly 16 U.S. states. This allows voters to make a single selection of a political party that then applies to all contests in which a member of that party is a valid option. There is evidence that the SPV option contributes to higher error rates on ballots [Campbell and Byrne 2009b, Herrnson et al. 2012] and it is not at all clear how SPV would interact with the direct access navigation model, as we did not provide an SPV option to voters in this experiment.

Given our result that the direct access DRE results in significantly more intentional abstention, one might argue this navigation style is not ideal, and that encouraging abstention is

not a desirable quality for a voting machine. The argument presented here is that we should ideally separate intentional abstention from unintentional undervoting, as including the former in residual error rates categorizes an intentional behavior with an error. What is an error is a normative argument contingent on the goals for our democratic instruments, and a highly debatable one. While some empirical work has examined hypothetical scenarios of election outcomes under full (hypothetically forced) turnout [e.g., Citrin et al. 2003], particularly Senate outcomes and not further down-ballot races, whether encouraging rational abstention should be a goal of an election system is a normative question we leave aside.

It is not immediately clear how this situation could be improved for any future direct access DRE. More visually salient highlighting of undervotes, as in [Campbell and Byrne, 2009a], might serve to better alert voters that some races have not been filled in. However, since many voters do abstain, particularly in down-race ballots, the issue is not simply whether there are any undervoted races, but whether only the intended abstentions are missing. Additional prompting of the voter when they attempt to cast the ballot (e.g., a warning message such as "There are contests for which you have made no selection. Are you sure you want to submit your ballot?") seems similarly ineffective exactly because some level of intentional abstention is not unusual. Sequential ballot presentation is likely a better solution.

This raises other interesting questions about what can or should be done to reduce unintentional abstention in other ballot formats. There is little that could be done with paper ballots (or punch cards), but sequential DREs could be programmed to not accept a ballot that did not have all races marked. In order to allow intentional abstentions, an explicit "abstain" option would have to be added to each contest. This style of ballot would likely run into legal problems in many U.S. jurisdictions. While as far as we know, no data on usability for such a system has ever been published, such a configuration could be quite onerous for voters, particularly those voters faced with many races who intend to abstain in most of the contests.

As noted in other research, participants were highly satisfied with their experience using the sequential DRE. Despite receiving significantly higher subjective usability ratings than any other method, there were no corresponding objective benefits to using the sequential DRE. Although voters preferred it, they were neither faster nor more accurate with it than with any of the older technologies.

Another issue that warrants closer scrutiny is the utility of the residual vote as a measure of accuracy. Consistent with [Campbell and Byrne 2009a], our data indicate that the residual vote and the true error rate may not be as tightly coupled as has traditionally been assumed. In both this experiment and Campbell and Byrne, the mean error rate generated by both measures was approximately the same. So, for use as a tool to detect large-scale problems across many voter, the residual error rate is indeed probably a good tool. It appears that, on average, the overall rate of wrong choice errors (included in the true error rate but not the residual vote rate) somewhat mirrors the intentional abstention rate (included in the residual vote rate but not the true error rate). There are probably circumstances where this breaks down, but as a low-precision tool, the use of residual vote rate is probably not a serious problem.

However, the near-zero correlation observed between residual vote rate and true error rate at the ballot level is cause for some concern. This suggests that the residual vote is incapable of identifying individual ballots, or probably even subsets of ballots, that contain errors. While wrong choice errors and intentional abstentions balance out in the aggregate, they do not do so at the level of individual ballots. Further research on the exact nature of this relationship is clearly warranted. Of course, this requires further laboratory research since the true error rate cannot be known in a real election for privacy reasons. Voting studies run in the laboratory environment, such as those reported here, complement field research and can shed light on issues that are impossible to address during real elections, such as the accuracy of the residual vote and evaluation of experimental voting system designs. The laboratory setting allows for more granular measurement of the objective and subjective facets of usability, both of which are important for meaningful comparisons of voting methods over time. Regardless of whether voting is electronic or paper-based, we must first understand the design space of our current systems before we can predict and evaluate the usability of emerging and future voting methods.

ACKNOWLEDGMENTS

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

Ansolabehere, S. and Stewart, C. 2005. Residual Votes Attributable to Technology. *The Journal of Politics*. 67, 2, 365–389.

Bailey, R. W. 1993 Performance vs. preference. I*n Proceedings of the Human Factors and Ergonomics Society 37th Annual Meet*ing, Human Factors and Ergonomics Society.

Brooke, J. 1996 SUS: A "quick and dirty" usability scale. I*n Usability Evaluation in Indust*ry, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McClelland, (Eds.). New York: Taylor and Francis.

Byrne, M. D., Greene, K. K., and Everett, S. P. 2007 Usability of voting systems: Baseline data for paper, punch cards, and lever machines. I*n Human Factors in Computing Systems: Proceedings of CHI 20*07, ACM.

Campbell, B. A. (2013). The Usability Implications of Long Ballot Content for Paper, Electronic, and Mobile Voting Systems. Doctoral dissertation, Rice University, Houston, TX.

Campbell, B. A. and Byrne, M. D. 2009a Now do voters notice review screen anomalies? A look at voting system usability. I*n 2009 USENIX/ACCURATE Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOT*E), Montreal, Canada.

Campbell, B. A. and Byrne, M. D. 2009b. Straight party voting: What do voters think*? IEEE Transactions on Information Forensics and Securi*ty. 4, 4, 718-728.

Citrin, J., Schickler, E., and Sides, J. 2003. What if everyone voted? Simulating the impact of increased turnout in Senate elections. *American Journal of Political Scien*ce. 47, 1, 75-90.

Everett, S. P., Byrne, M. D., and Greene, K. K. 2006 Measuring the usability of paper ballots: Efficiency, effectiveness, and satisfaction. I*n Proceedings of the Human Factors and Ergonomics Society 50th Annual Meet*ing, Human Factors and Ergonomics Society.

Everett, S. P., Greene, K. K., Byrne, M. D., Wallach, D. S., Derr, K., Sandler, D., and Torous, T. 2008 Electronic voting machines versus traditional methods: Improved preference, similar performance. I*n Human Factors in Computing Systems: Proceedings of CHI 20*08, ACM.

Felten, E. 2009. Finnish Court Orders Re-Vote After E-Voting Snafu. https://freedom-to-tinker.com/blog/felten/finnish-court-orders-re-vote-after-e-voting-snafu/

Greene, K. K., Byrne, M. D., and Everett, S. P. 2006. A comparison of usability between voting methods. *2006 USENIX/ACCURATE Electronic Voting Technology Worksh*op.

Herrnson, P. S., Hanmer, M. J., and Niemi, R. G. 2012. The impact of ballot type on voter errors. *American Journal of Political Scien*ce. 56, 3, 716–730.

Herrnson, P. S., Niemi, R. G., Hanmer, M. J., Bederson, B. B., Conrad, F. C., and Traugott, M. W. 200*8 Voting technology: The not-so-simple act of casting a ballo*t. Brookings Institution Press.

Kimball, D. C. and Kropf, M. 2008. Voting Technology, Ballot Measures, and Residual Votes. *American Politics Researc*h. 36, 479–509.

Kimball, D. C., Owens, C., and McAndrew, K. 2001. Who's Afraid of an Undervote*? Annual meeting of the Southern Political Science Associatio*n, Atlanta, GA.

Laskowski, S. J., Autry, M., Cugini, J., Killam, W., and Yen, J. 2004. *Improving the Usability and Accessibility of Voting Systems and Produc*ts. NIST Special Publication.

Mebane, W. R. 2004. The wrong man is president! Overvotes in the 2000 presidential election in Florida. *Perspectives on Politics*. 2, 3, 525–535.

Nichols, S. M. and Strizek, G. A. 1995. Electronic Voting Machines and Ballot Roll-off. *American Politics Quarterl*y. 23, 300–318.

Nielsen, J. and Levy, J. 1995. Measuring usability: Preference vs. performance. *Communications of the A*CM. 37, 4, 66–75.

Norden, L., Kimball, D., Quesenbery, W., and Chen, M. 200*8 Better ballo*ts. Brennan Center for Justice at NYU School of Law.

Ohio Secretary of State. 2007. Project EVEREST: Risk Assessment Study of Ohio Voting Systems Executive Report. http://www.sos.state.oh.us/SOS/upload/everest/00-SecretarysEVERESTExecutiveReport.pdf

Ryan, P. Y. A., Bismark, D., Heather, J., Schneider, S., Xia, Z. 2009. The Prêt à Voter verifiable election system. *IEEE Transactions on Information Forensics and Security, 4,* 662–673.

Stein, R. M., Vonnahme, G., Byrne, M., and Wallach, D. 2008. Voting technology, election administration, and voter performance. *Election Law Journ*al. 7, 123–135.

Stewart, C. 2006. Residual Vote in the 2004 Election. *Election Law Journ*al. 5, 158–169.

Traugott, M. W., Hanmer, M. J., Park, W.-h., Herrnson, P. S., Niemi, R. G., Bederson, B. B., and Conrad, F. G. 2005. The impact of voting systems on residual votes, incomplete ballots, and other measures of voting behavior. *Annual conference of the Midwest Political Science Associatio*n, Chicago, IL,. 7–10.

Wand, J. N., Shotts, K. W., Sekhon, J. S., Mebane, W. R., Herron, M. C., and Brady, H. E. 2001. The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida. *American Political Science Review*. 95, 4, 793–810.