Now Do Voters Notice Review Screen Anomalies? A Look at Voting System Usability

Bryan A. Campbell, Michael D. Byrne

Department of Psychology
Rice University
6100 Main Street, MS-25
Houston, TX, 77005-1892, USA
Bryan.Campbell@rice.edu, Byrne@rice.edu

Abstract

Everett (2007) showed that about two thirds of voters do not notice when the review screen on an electronic voting machine does not agree with the selections they intended. However, the instructions given to voters in those experiments did not emphasize the need for verification and the design of the review screen did little to aid voters in detecting anomalies. This research follows up that work, but this time with improved instructions and with a re-designed review screen that makes certain conditions, particularly undervotes, more visually salient. This did increase the detection rate, but only up to 50%. Our findings also extend Everett's research; in general, people tend to prefer electronic voting machines over other technologies such as punch cards, lever machines, and even paper ballots. This was again true, but only for voters who failed to notice any anomaly. In addition, voters took longer to vote with the DRE than with older technologies, but with no concomitant decrease in error rate. We also show that the relationship between true error rates and the oft-used residual vote rate is not straightforward.

1. Introduction

Security is obviously a critical concern in any election, particularly those involving voting computers, more commonly called DREs (for direct recording electronic). However, as we have argued in the past (e.g., Byrne, et al., 2007; Everett, et al., 2008), usability is also a critical concern when it comes to voting systems.

These two areas meet when considering the role of the voter in ensuring the integrity of their own ballot. For instance, in the wake of legitimate security concerns, many people have advocated the adoption of having DREs produce redundant paper records that the voter can confirm matches their intent; these are also known as VVPATs (voter-verifiable paper audit trails). For VVPATs to be effective in detecting malicious or malfunctioning DREs, the voters must verify that the paper record is correct. However, it is not known how accurate voters are at doing this, nor how much time this adds to the voting process; these are essentially usability concerns.

However, even paperless DREs offer a first line of defense in the form of a review screen. Unfortunately, Everett (2007) reported two experiments that showed that nearly two-thirds of voters did not notice when the review screen did not match the selections they had made. This is a distressing result, but Everett's study had some limitations. In particular, voters were not specifically instructed on the importance of review screen verification, and the review screen itself did very little to aid voters in detecting anomalies. Undervotes (that is, races with no selection made) were not highlighted, and the review screen did not include party information, which several voters indicated they would have used had it been present.

The research reported here is a replication of Everett (2007), but with those issues corrected. Before getting into the details of the current experiment, it is important to consider the Everett (2007) research in more detail as well as provide some additional background on voting system usability.

1.1 Previous Research on Detection of Review Screen Anomalies

There are a variety of reasons why it is important for voters to carefully check the review screen on a DRE. Review screens provide the opportunity to check for errors, which could be due to voter mistakes but could also be the result of faulty DRE hardware or software, such as a mis-calibrated touchscreen. A difference between the voter's intent and the review screen could also be the result of malicious software which changes voter's selections before the ballot is cast. This would obviously not be the most sophisticated attack; a better attack would change votes without also changing the review screen. Regardless, review screens potentially provide at least some protection from malicious or malfunctioning DREs.

However, for this to be an effective method of discovering a malfunctioning DRE, voters must actually check the review screen and do so accurately. But do they? This was the question Everett (2007) set out to answer. Two experiments were conducted in which the review screens did not match what voters had selected previously.

The critical finding was that across the two studies, the majority of the participants did *not* notice the mismatch. In the first experiment (reported as Study 2), two or eight entire contests were added (contests which had not been previously seen) or subtracted (contests that had been previously seen). In the second (reported as Study 3), one, two, or eight contest changes were made to the review screen. These changes were either to an opposing candidate (e.g., a Democratic candidate to a Republican candidate) or to an undervote (i.e., an absence of a vote in the Presidential race) and appeared in either the first half or second half of the 27-contest ballot. In Study 2, only 32% of participants noticed anomalies on their review screens. In Study 3, only 37% of participants noticed anomalies on their review screens. This means that roughly two-thirds of participants did *not* notice up to 8 anomalies on their review screen.

Beyond the fact that most voters did not notice, there were factors that were systematically related to whether or not voters noticed. In both experiments there were two critical and statistically reliable predictors of noticing review screen anomalies:

Time voters spent on the review screen. The longer a voter spent on the review screen, the more likely they were to notice the anomaly. It is not clear whether this is because spending more time on the review screen increased the chance of catching the anomalies, or whether noticing an anomaly caused voters to spend more time examining the review screen.

The "information condition" a voter was in. Half the voters were given a piece of paper which directed them to vote for specific candidates, while half were given a voter's guide and allowed to make their own selections. Those who were directed were more likely to notice an anomaly. This is important because usability requirements for voting systems, such as those in the draft of the next voluntary voting system guidelines (VVSG) specify that user testing will be done using the directed approach. Thus, they are likely to underestimate the scope of this problem.

There were other factors which appeared in one experiment and not the other and were statistically smaller than the other two predictors. In Study 2, a personality variable, "openness to experience" was related to noticing. As described by McCrae and Costa (1987), those participants who receive higher scores in openness to experience tend to be "creative" and "curious" individuals. Participants with these traits noticed review screen anomalies slightly more often. In Everett's Study 3, age was also negatively related to detection; that is, older voters were somewhat less likely to notice. Also in Study 3, voters who self-reported that they followed reports of election security issues in the news were more likely to notice. These effects were not nearly as strong as the effects of time and information condition.

Equally important are the variables which were *not* meaningful predictors of anomaly detection. The number of races altered was not a factor in noticing; voters were almost as likely to notice when one race had been changed as they were to notice that eight races had been changed. Ballot position was also not a factor; top-ballot changes were not reliably more likely to be noticed than changes further down the

ballot. Finally, the noticing rates were very similar across the two experiments, suggesting that the type of change was also not a factor. These are surprising results; intuition suggests that adding 8 additional races to a 27-race ballot should be more noticeable than changing a single race from one candidate to another. However, the data do not support such a story; time and information condition were the only consistently reliable predictors.

However, there are a few caveats regarding Everett's findings. One issue is that participants were never explicitly told to check the review screen. It could be that simply instructing voters to utilize the review screen may have increased the rate of noticing anomalies. At the time, DREs were still relatively new to the region in which the research was conducted and most voters only get use the technology a few times a year at most.

Another concern is that the user interface of the experimental DRE did little to make it easy for voters to perform their checks. This review screen is presented in Figure 1. The problem may have been that there was too little information on the review screen such that the cost of verification may have been considered too high by the voters.

Perhaps, then, detection performance might be improved by altering instructions to emphasize the importance of verifying the review screen contents and changing the user interface in the review screen; this is precisely the motivation for the present study.

Both Everett's previous work and the current work is in the context of research on voting system usability, and so some background on this is also important.

1.2 Voting System Usability

Usability is not an area which received much attention in the conduct of elections until the now-infamous "butterfly ballot" in the 2000 election in Florida. However, this is a critical factor in election integrity. How can one be confident that the results of an election with a 1% margin of victory actually reflect the will of the electorate if 2% or more of the voters made errors in casting their ballots?

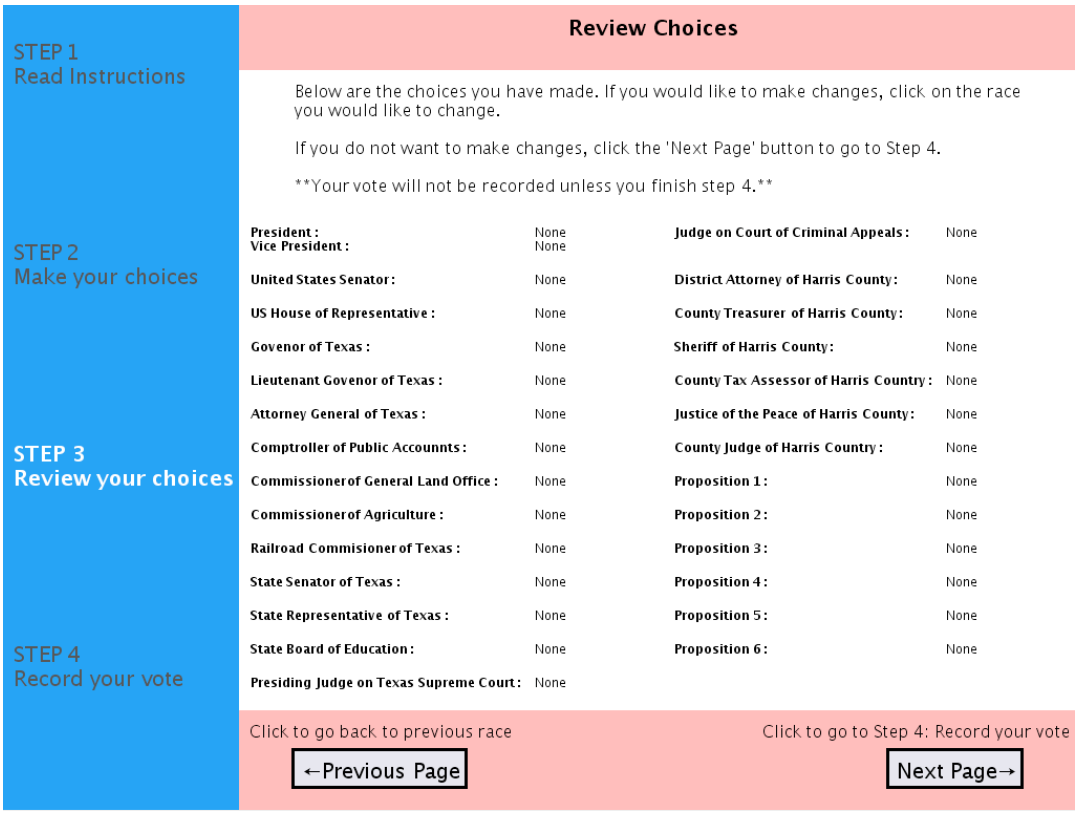

Figure 1. Review screen used in Everett (2007).

Consider the amount of time and energy which has been spent on issues of voting system security (e.g., the California Top-to-Bottom review; the Ohio EVEREST review; research like Rubin, 2006; Sandler, et al., 2008) when there is, as yet, no conclusive evidence that any election has even been electronically stolen. Usability, on the other hand, has almost certainly played a role in determining the outcome of multiple elections (Brennan Center, Norden, et al., 2008), including the 2000 U. S. Presidential election (Mebane, 2004; Wand, et al. 2001). Consider the Brennan Center report. In nearly every case they detailed, the residual vote rate was higher than the margin of victory. A residual vote rate is the difference between the number of eligible voters who cast ballots at the polls and the number of valid ballots received. A residual vote is the result of an omission (which may or may not be an actual error; voters can abstain if they so desire), which is termed an *undervote*, or some kind of spoiled vote such as making two selections in a race where voting for only one candidate is allowed (an *overvote*).

Government involvement in usability in the United States was initiated with and continues under the purview of the Help America Vote Act (HAVA – United States Government, 2002) and the Elections Assistance Commission (EAC) created as a result of HAVA's Congressional passage. In addition to the creation of the EAC, HAVA sanctioned a joint effort between the EAC and the National Institute of Standards and Technology (NIST) to review and make recommendations for improving the usability of current voting technologies.

In addition, by mandate of HAVA, the EAC and NIST were required to produce a report that:

> "…assesses the areas of human factors research, including usability engineering and human-computer and human-machine interaction, which feasibly could be applied to voting products and systems, including methods to improve access for individuals with disabilities (including blindness) and individuals with limited proficiency in the English language and to reduce voter error and the number of spoiled ballots in elections" (United States Government, 2002).

NIST Special Publication 500-256 (Laskowski, et al., 2004) is the product of the HAVA mandate. Within this special publication, Laskowski et al. (2004) describe three usability metrics, borrowed from ISO standard 9241-11 regarding general usability, used to define usability as it pertains to voting technologies (ISO 9241-11, 1998). The first of these usability metrics is effectiveness. Effectiveness is described within the NIST report as "the accuracy and completeness" of a voting technology. The second usability metric is efficiency. Efficiency is described within the NIST report as "the resources expended by the user," generally in terms of time taken to vote. The third usability metric is satisfaction. Satisfaction is the only subjective usability metric represented in the NIST report and is described as voter "comfort and acceptability." Effectiveness can be measured in terms of ballot errors and corresponding error rates. Efficiency can measured in terms of time on task, or ballot completion times. Finally, satisfaction, the subjective usability metric, can be measured by administering a questionnaire such the System Usability Scale (or SUS; Brooke, 1996).

Since the publication of that report, several studies of voting system usability have been conducted that use some or all of the NIST metrics. Herrnson, et al. (2008) conducted a large-scale study of commercial DREs and found that there were substantial differences in error rates between different commercial DREs; all of them were fairly high, upwards of 2.5% per race. However, they did not measure efficiency and their metric for satisfaction was non-standard.

Another program of research using all three of the NIST metrics is ours, e.g., Greene, et al. (2006), Byrne, et al., (2007), Everett, et al., (2008). This research has systematically compared various older voting technologies on all three metrics as well as comparing them with the VoteBox DRE. Voting method appears to make little difference in efficiency. They make a small difference in effectiveness with punch cards and lever machines sometimes faring worse than other technologies. However, the difference in satisfaction is substantial; people generally prefer paper ballots to punch cards and lever machines, and they prefer VoteBox over all other technologies, sometimes quite strongly. The disconnect between the subjective scale and objective performance is somewhat distressing; VoteBox is strongly preferred despite a clear lack of objective advantage, particularly when compared with paper.

This suggests some backlash in cases where DREs are abandoned in favor of technologies perceived to be more secure, such as paper.

The objectives of the current experiment are twofold. First, can the rate of anomaly detection be improved by changes to instructions and the review screen? Second, we want to attempt to replicate previous findings on usability of DREs vs. traditional technologies.

2. Method

2.1 Participants

We recruited 108 participants from the greater Houston area via an advertisement placed in a major local newspaper. Of the 108 participants (M_{Age} = 43.1 years, *SD* = 17.9 years), 60 were female and 48 male. The only requirements for participation were (1) native English speaking and (2) age of 18 years or older (i.e., eligible to vote). Not all of our participants divulged their voting histories; *N*'s are reported where appropriate. Participants were paid $25 for their time and payment was not a function of voting performance. Voting performance in the laboratory context has been shown not to be affected by performance-based monetary incentive (Goggin, 2008). All but 7 participants reported having normal or corrected to normal vision.

Twenty-three participants reported having voted in 10 or more previous national elections while 21 reported having voted in 10 or more non-national (e.g., state and local) elections. The average number of previous national elections was 5.8 (*SD* = 7.3, *N* = 98) and the average number of previous non-national elections was 6.3 (*SD* = 11.2, *N* = 81). In addition, we observed a substantial range of incomes as well. Finally, on a 10-point Likert scale, the mean level of self-rated computer expertise was 6.2 (*SD* = 2.5) with 59 participants reporting 0 to 20 hours of computer use per week (the remaining 49 reported greater than 20 hours per week). Additional demographics are shown in Table 1.

Table 1. Participant demographics. One participant failed to divulge their education and ethnicity.

Level of Education	Frequency	Percentage
High school or less	10	9.3%
Some college	42	39.3%
Bachelor's degree or equivalent	40	37.0%
Postgraduate degree	15	14.0%
Race / Ethnicity	**Frequency**	**Percentage**
Caucasian	51	47.7%
African American	36	33.6%
Other	20	18.7%

2.2 Design

The design was complex with a substantial number of independent and dependent variables. The manipulated independent variables were:

Voting system (2 levels, within subjects): All participants voted twice; once on the DRE and once with another voting technology. Participants were instructed to vote for the same candidates both times and technology order was counterbalanced. There was no non-anomalous condition in our design as previous research has investigated performance using the DRE and other voting technologies in the absence of anomalies (see Byrne et al., 2007; Everett, 2007; Everett et al., 2006, Everett et al., 2008; Goggin, 2008; & Greene et al., 2006).

Other voting technology (3 levels, between subjects): Participants' non-DRE voting system was either bubble-style paper ballots, punch cards, or lever machines. Any change in voting technology should be at least as good as the technology it replaces. As of the 2008 Presidential election, all 3 other voting technologies were in use in the United States (Brace, 2008). In order to make meaningful comparisons concerning the usability of DREs, it is necessary to include such older voting technologies.

Information condition (3 levels, between subjects): One third of the participants were given a voter's guide (modeled after the League of Women Voters product) and told to vote however they wished; this is termed the "undirected" condition. After voting in the in the undirected condition, participants were given an exit interview in order to ascertain for whom they had voted. The exit interview allowed us to determine errors for participants in this condition. In the other two groups, participants received a sheet of paper

instructing them to vote for specific candidates. In the "fully directed" condition, this sheet contained a selection for every race. In the "directed with roll-off" condition, the sheet directed participants to omit some races. The probability of a race being omitted increased further down the ballot to mirror real-world abstention patterns.

When participants voted with the DRE, the review screen did not match their selections. There were three variables manipulated:

Anomaly type (2 levels, between subjects): This concerns how votes were changed on the review screen. For half the participants, votes were changed to a different candidate, for the other half, votes were replaced with undervotes (that is, the selection was marked as "none" for the changed race).

Number of anomalies (3 levels, between subjects): Either 1, 2, or 8 selections were altered on the review screen.

Anomaly location (2 levels, between subjects): For half the participants, the anomalies were in the first 14 races on the ballot, for the other half of the participants, changes were in the last 13 races.

The three variables involving anomaly characteristics were fully crossed whereas the other two were not (though participants were still randomly assigned in all conditions).

Demographic variables of age and education were also included as independent variables in all analyses. On the dependent side, there were also multiple variables:

Detection, also termed "noticing." Did the participant notice any anomaly on the review screen? We used the generous criterion used in Everett (2007): we directly asked participants if they noticed any discrepancies on the review screen. This is obviously an imperfect self-report measure, but even by this measure a strong majority of participants in Everett's research failed to report noticing.

Efficiency. This was the time taken to vote, measured with a stopwatch from when the participant started their ballot until it was complete.

Satisfaction. This was measured with the SUS (Brooke, 1996), which is a 10-question instrument using Likert scales.

Effectiveness. This was measured by looking at error rates. Obviously, this requires a concrete definition of "error." In the two directed conditions, this is straightforward: did the participant vote as instructed? In the undirected condition, participants were interviewed after voting and asked to reveal their choices. Thus, with two votes cast and the exit interview, errors could be defined by majority rule: any response that did not match the other two was considered an error.

Errors were further subdivided into multiple types. *Overvote errors* occurred when participants selected more than one candidate in a single-candidate race. Note that both the DRE and the lever machine prevent this kind of error. *Undervote errors* occurred when voters omitted a race in which they should have cast a vote. *Wrong choice errors* occurred when participants selected a candidate different from the one intended. *Extra vote errors* occurred when participants cast a vote in a race they meant to omit.

We also tracked the rate of intentional undervotes; that is, how often participants omitted races intentionally. This measure is only meaningful in the undirected condition, since the rate of intentional undervoting is zero in the fully directed condition and is set by the list in the other conditions.

This allows us to compute what would be reported in an actual election as the *residual vote rate*, which is the sum of overvote errors, undervote errors, and intentional undervotes. The true error rate, of course, also includes wrong choice errors and extra vote errors. This allows us to compare the residual vote rate and the true error rate for the undirected condition.

Finally, we also asked participants how useful they thought the review screen was to them, and asked them how carefully they checked the review screen,

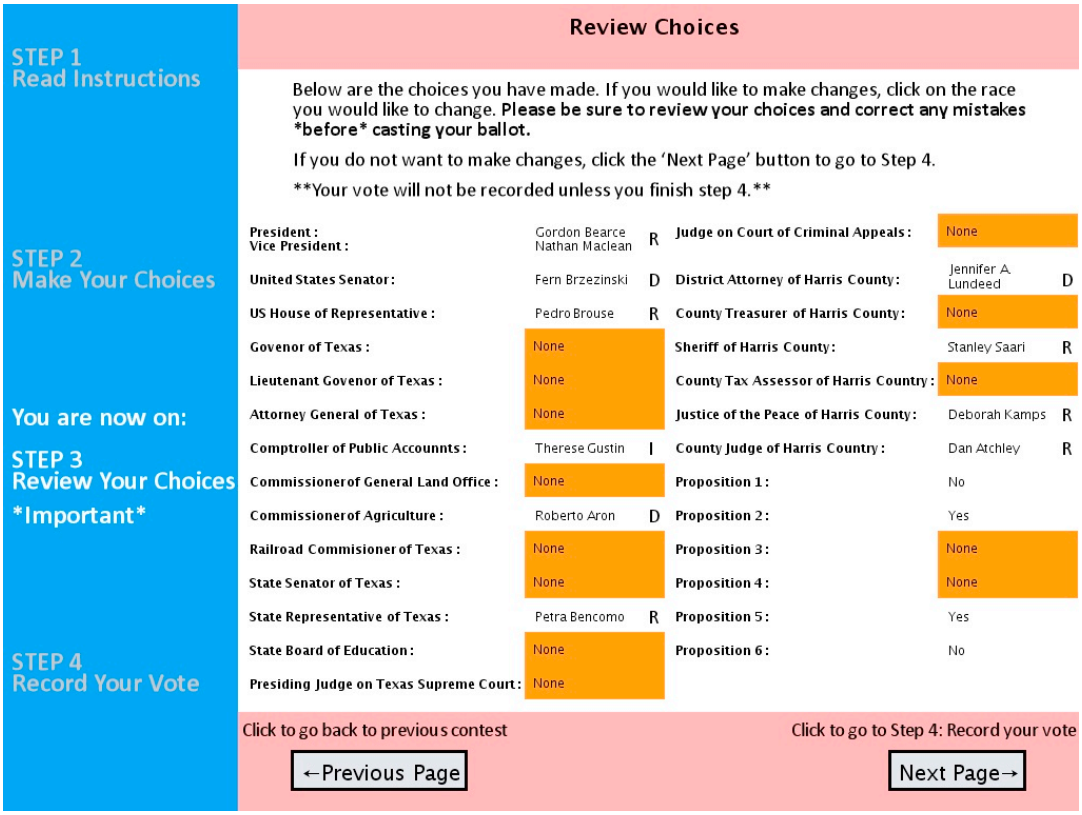

Figure 2. Updated review screen used in this experiment.

either "not at all," "somewhat carefully," or "very carefully."

2.3 Materials

The most critical difference between this experiment and the ones presented in Everett (2007) are here. In particular, participants were instructed that verification was an important step. In addition, the actual review screen on the VoteBox DRE was altered. We made two changes: party information was included, and undervotes were highlighted with a distinctive red-orange color. See Figure 2 for the current review screen.

VoteBox (Sandler, et al., 2008) is a Java-based DRE platform that includes special code to support usability experiments. First and foremost for present purposes, it supports the necessary modifications to the review screen. In addition, it collects, time stamps, and logs all user actions (e.g., mouse clicks).

The paper ballots were standard bubble-type ballots, in the style commonly used for optical scan machines. The punch card stations were VotoMatic IIIs purchased at an auction from Brazoria County, TX. To vote, participants turn the pages to reveal candidates in each race and use a metal stylus to punch holes corresponding to candidate of their choice. The lever machines were manufactured by Automatic Voting Machines, Inc. and purchased from an auction in Victoria County, TX. To begin voting, participants pull a red handle to the right side. Candidate selections were made by pressing the levers down so that an arrow pointed at the chosen name.

The ballots used in prior research (Byrne et al., 2007; Everett et al., 2007; Greene et al., 2006) as well as the ballots used in the current research were identical. Our ballot consisted of 27 contests featuring 21 candidate races and 6 propositions. Candidate names were randomly generated as this has been shown not to have an effect on voting performance (Everett et al., 2006; Greene et al., 2006) whilst alleviating concerns about forcing participants to reveal any actual past or future votes. The 21 candidate races and 6 propositions, though fictional, were representative of contests that have appeared in

recent greater Houston area elections. In our ballot, there was no option to cast a straight party or straight ticket vote (as is currently allowed in Texas). As noted in previous research (Byrne et al., 2007), in the U.S., laws on straight party voting vary from state to state, and in some instances, by county. It would have been difficult to include all or many variations of the straight party voting provision in our current design, so our ballot did not include the option to cast a straight party vote. There is also research that suggests that straight-party voting can be confusing to voters (Redish, 2008) and we did not want such confusion to contaminate our results.

Participants who received a slate of candidates to vote for were given slates that contained 85% Democratic candidates, 85% Republican candidates, or an equal mixture of Democratic and Republican candidates. Use of fictional names with real political parties is an explicit effort to maintain as much realism as possible without compromising participants' privacy; revealing party preference does not reveal actual choices in real elections.

2.4 Procedure

After informed consent and other administrative details, participants were given a either a list of candidates to vote for (full or with roll-off) or a voter guide. They were also given a set of voting instructions, and time to read, understand, and ask questions about each. When all was understood, participants were directed to a voting station and allowed to complete their first ballot to the best of their ability. Upon completion of the first ballot, participants were directed to another voting station and allowed to complete their second ballot to the best of their ability. Participants were instructed, both orally and in writing, to vote exactly the same way on both ballots prior to any voting. Immediately following voting on each technology, participants were given the SUS regarding the voting system they had just used. Participants who were given a voter guide were then given an exit interview to ascertain who they had voted for. This exit interview allowed us to determine errors and error rates for participants in the voter guide condition by severing as a participant-generated slate. Ballot completion times for both voting sessions (DRE and Other) were captured via stopwatch, as this was the most effective method for capturing completion time data for the non-electronic voting methods. After both voting sessions were complete, participants were given a demographics and voting experience questionnaire. Upon completion of this questionnaire, participants were debriefed and paid for their time.

3. Results

3.1 Detection of Review Screen Anomalies

The primary result of interest is how many people noticed the mismatch between the review screen and their previous selection(s). Recall that in Everett (2007), approximately one-third of voters noticed. Did the changes that were specifically designed to increase noticing improve detection performance?

It appears they did; exactly 50% of the voters noticed a review screen anomaly; the 95% confidence interval ran from 40.1% to 59.9%. This is clearly an improvement over Everett's results. However, the improvement is somewhat modest; this is still a far cry from having almost all voters noticing.

Were the same factors at play here as found by Everett? Recall that time spent on the review screen and information condition were the major predictors in her experiments.

Three logistic regressions were conducted: one for situational variables (that is, those experimentally manipulated), one for personal characteristics such as age, computer experience, education, and personality variables, and one for performance factors including time on the review screen as well as self-reported care in examining the review screen and self-reported rating of review screen usefulness.

The type of anomaly was a factor: 61% of the participants who had their votes changed to undervotes noticed, whereas only 39% of those who had their votes changed to another candidate noticed. This difference was statistically reliable ($w = 5.77$, $p = .02$). This suggests that the addition of the color marking of undervotes did result in an improvement in detection performance.

Unlike with Everett's study, information condition was not as powerful a variable; see Table 2 for the

detection rate in the three conditions. The generalized difference between all three conditions is suggestive but not significant at conventional levels, $w = 4.61$, $p = .10$. However, a specific contrast comparing the fully directed condition against the mean of the other two conditions is significant, $w = 4.55$, $p = .03$. This is also consistent with the idea that it is undervotes which drive noticing, since in the fully directed condition, the review screen should contain no undervotes.

Table 2. Detection rate by information condition.

	Fully Directed	Undirected	Directed with roll-off
Detection rate	64%	44%	42%

Table 3 presents the detection rate as a function of number of anomalies; again, the results were suggestive but not significant at conventional levels, $w = 4.61$, $p = .10$.

Table 3. Detection rate by number of anomalies.

	1 anomaly	2 anomalies	8 anomalies
Detection rate	36%	56%	58%

Top-ballot races were noticed at at rate of 57% while lower-ballot races were noticed at a rate of 43%. This difference is also suggestive but not significant, $w = 2.68$, $p = .10$.

Noted earlier, an option to cast a straight party vote was not present on our ballot, However, in the undirected condition, participants could still vote a straight party ticket by selecting all the candidates of a particular party. Four participants did just this, although it did not have an impact on anomaly detection. There were also no effects of any personal variables such as age, education, computer experience, news following, or personality variables.

Finally, we examined the effect of review-related variables, including time spent on the review screen as well as self-reported care in examining the review screen and self-reported rating of review screen usefulness.

Time spent on the review screen was again a very strong predictor of detection, with more time spent on detection associated with a higher rate of detection, $w = 9.11$, $p = .003$. Those who did not notice spent an average of 40 sec on the review screen, whereas those who noticed spent an average of 130 sec on the review screen. Again, we cannot be sure about the direction of causality here; perhaps those who spend more time do so because they detected an anomaly.

Table 4 shows the detection rate as a function of self-reported care in reviewing the summary screen; this effect was also statistically reliable, $w = 6.24$, $p = .04$. Again, causal direction is not entirely clear here, as people may have reported that they were careful because they noticed. Furthermore, 28% self-reported that they were very careful, and yet they missed the discrepancy.

Table 4. Detection rate by self-reported care in reviewing the summary screen.

	How carefully did you review the summary screen?		
	Not at all	**Somewhat**	**Very Carefully**
Detection rate	0%	13%	72%

Overall, these results, although not entirely conclusive, suggest that the changes in instructions and review screen not only improved detection performance but increased context sensitivity. That is, Everett found no effect of number of changes or ballot position but we see hints of such effects here.

3.2 Satisfaction

Data for three subjects was incomplete for the SUS, so those subjects were excluded from the analysis. We also included whether or not participants noticed an anomaly as an independent variable. As it turned out, this had an impact on participants' rating of the DRE. Figure 3a presents mean SUS ratings for voters who noticed an anomaly and Figure 3b for those who did not. The difference here is quite dramatic: the preference for DRE vs. bubble and for punch card

was reversed as a function of noticing; overall this interaction was statistically reliable, $F(2, 43) = 4.24$, $p = .02$. Overall, the DRE was preferred to other technologies for those who did not notice, but *not* for those who did; this interaction was also statistically reliable, $F(1, 43) = 8.43$, $p = .006$.

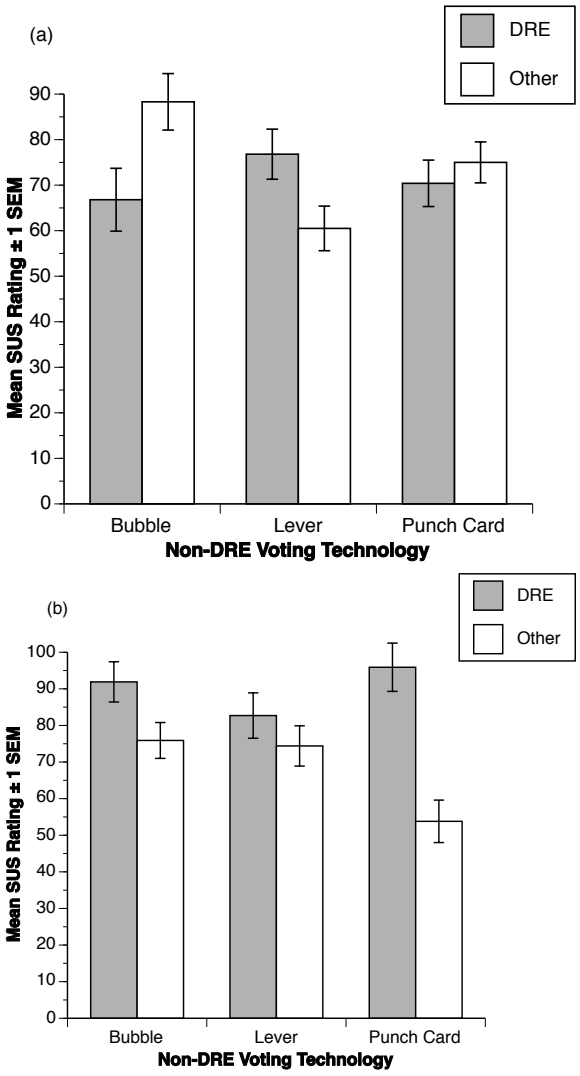

Figure 3. Satisfaction as measured by SUS as a function of voting technology. (a) shows the results for voters who noticed an anomaly, and (b) shows the results for voters who did not.

Everett (2007) did not include noticing in the analysis of SUS scores, so this is a new result. In some sense this is not surprising—noting that the DRE was "cheating" caused voters to like it less—but the magnitude of this effect would have been difficult to predict.

Another new effect was the interaction of age and voting technology, displayed in Figure 4. Older voters rated both DREs and non-DREs just as high, but younger voters clearly favored the DRE. This effect was statistically reliable, $F(1, 43) = 8.65$, $p = .005$.

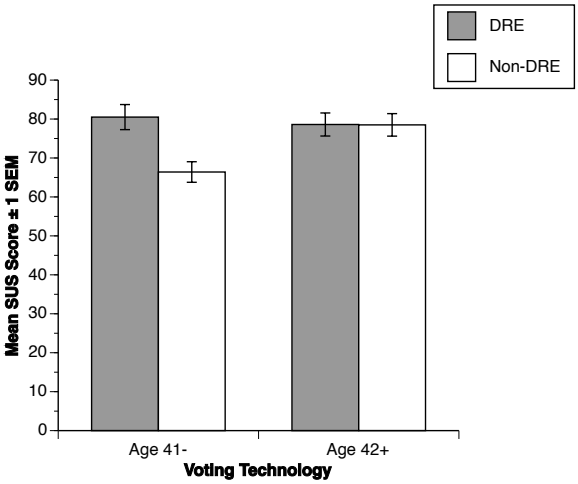

Figure 4. Mean SUS score as a function of age and voting technology.

3.3 Effectiveness

Ten participants were removed from this analysis due to having error rates greater than 15% on both voting methods as this suggests that those participants did not have an adequate understanding of the task instructions. (For instance, one subject in one of the directed conditions refused to give up his personal party allegiance and voted almost entirely for his favored party despite instructions to follow the list given to him). Although we cannot be absolutely certain, at least 5 participants appeared to exhibit this behavior. Additionally, in the undirected condition multiple participants appeared unable or unwilling to recall who they voted for in the exit interview we administered. These participants were subsequently removed from the analysis due to their high error rates.

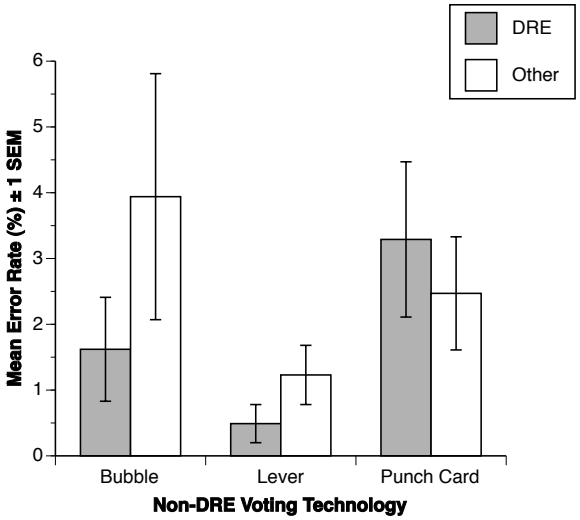

Figure 5. Mean overall error rate as a function of voting technology across all participants.

Overall error rate as a function of voting system are shown in Figure 5. As can be seen from the graph, the data are fairly noisy; none of the independent variables, including whether or not participants noticed the anomalies or not, had a statistically reliable impact on overall error rate,

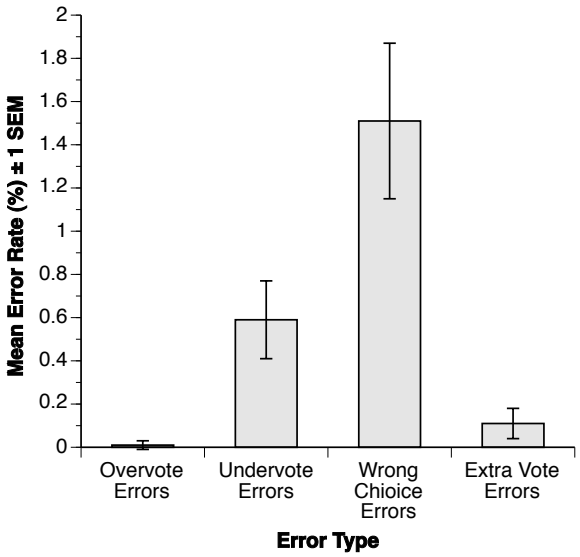

Figure 6. Mean error rate by error type.

In fact, the only effect of note here was the difference in rate for the four kinds of errors as seen in Figure 6. Wrong choice errors were the most common type of error. The difference between error types was reliable, $F(3, 291) = 12.21$, $p < .001$.

Another interesting question concerns the relationship between the true error rate and the residual vote rate. How well do these measures correspond? At the aggregate level, the agreement is moderate. Figure 7 shows the mean values for both true error rate and residual vote rate. None of the differences there are statistically reliable.

On the other hand, we cannot conclude that the two measures are highly related at the individual level. For DREs, the correlation between the two measures is only $r(32) = .30$, $p = .095$ and for non-DRE, the correlation is a mere $r(32) = .02$, $p = .89$. This suggests that researchers should use extreme caution when using residual vote rate as a proxy for actual error rate; the two may not be as tightly related as many believe.

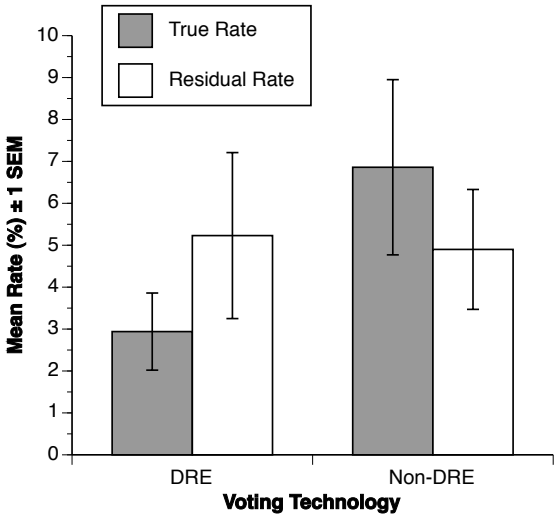

Figure 7. Mean true error rate vs. mean residual vote rate by voting technology.

3.4 Efficiency

Twelve participants were removed from the analysis for irregular voting times, defined as being more than 3 interquartile ranges above the 75th percentile[1].

[1] Most of the participants removed in this analysis were those who, after declining to read the voter guide provided in the undirected condition and then indicating their readiness to vote, reversed that decision and read the voter guide (which participants were allowed to keep with them) before and during voting, but after timing had already begun. Two participants removed from this analysis were the same as those removed in 3.3.

As can be seen in Figure 8, the DRE was consistently slower than the other technologies, $F(1, 37) = 26.80$, $p < .001$. We initially suspected that this was a result of voters noticing discrepancies on the review screen taking longer, and they did, on average, take about 80 seconds longer when using the DRE, but this difference was not quite statistically reliable, $t(94) = 1.83$, $p = .07$.

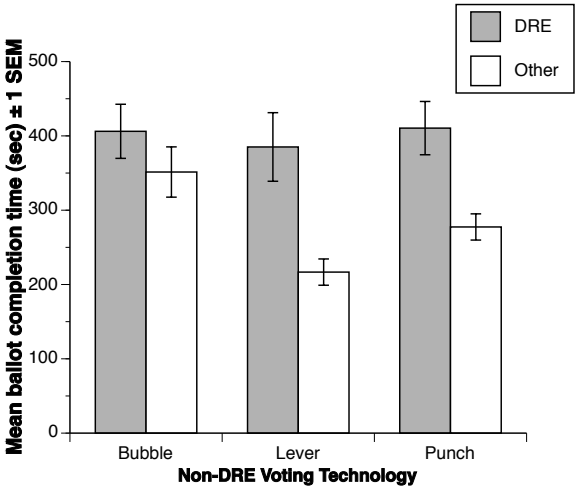

Figure 8. Ballot completion time as a function of voting technology.

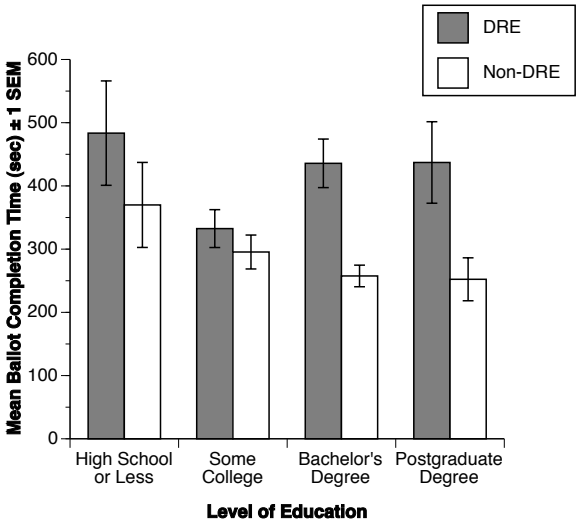

Figure 9. Ballot completion time as a function of education and voting technology.

There was also an interaction between education and DRE vs. non-DRE, as shown in Figure 9. There is no evidence for an effect of education on DRE time, $F(3, 37) = 1.31$, $p = .29$, but there is some evidence for an effect on non-DRE time, $F(3, 37) = 2.84$, $p = .05$. We had previously found some evidence for an effect of education on ballot completion time with non-DREs (Byrne, et al., 2007), so this result is not surprising. However, it is unclear why this appears not to apply to DREs.

There were some additional higher-order interactions, but these were not deemed meaningfully interpretable and so they have been omitted.

4. Discussion

Despite improvements to the user interface, only 50% of our participants noticed up to 8 review screen anomalies. This is an improvement over Everett (2007), particularly in the case of detecting changes when votes were replaced with undervotes. However, this suggests that simple GUI and instructional improvements may not be enough to drastically increase review screen anomaly detection.

This also raises questions about the utility of VVPATs. While it might be possible that people are more diligent about checking VVPATs because the purpose of the VVPAT is clear, we are skeptical in this regard; generally speaking, VVPATs are printed on inexpensive thermal printers and are thus are more difficult to read than the modern LCD displays used by most DREs. Our research is by no means conclusive with respect to VVPATs since we did not directly measure them, but these results certainly suggest caution when making assumptions about how many voters will actually detect discrepancies on a VVPAT. Note that more than a quarter of our voters claimed to have checked the review screen "very carefully" but yet failed to note any discrepancies.

We were able to successfully replicate numerous aspects of Everett's earlier work. In particular, the total time spent on the review screen was the best predictor of noticing the review screen anomalies. Additionally, in this case we also had some evidence that other factors like ballot position and number of anomalies may play a role.

This experiment also generated some important new findings. Previous research has shown the VoteBox DRE to be consistently preferred to other technologies (e.g., Everett, et al., 2008) despite

offering no advantages in terms of objective performance on time and errors. This was true for our participants, but only those who failed to detect any anomalies, despite the fact that performance in terms of time was actually *worse* with the DRE.

However, this did not apply to voters who did notice a discrepancy on the review screen. Those voters were, overall, much less enthralled with the DRE. This suggests that it may be possible to couple preference and performance if people have a negative experience with a voting technology.

Another interesting result concerns the relationship between the true error rate and the residual error rate. The residual vote rate is the *de facto* standard for post hoc evaluations of voting system effectiveness. (For example, it is used as the primary measure of voting system effectiveness in Norden, 2008). However, the residual vote rate has two shortcomings in this regard: first, it is based in part on undervotes, which may or may not be errors; second, it does not include any measure of wrong choice errors—that is, when a voter selects a candidate other than the one intended. (It is noteworthy that wrong choice errors were the most common errors made in this experiment). These shortcomings have long been acknowledged. However, since it is impossible in a real election to measure the true error rate, these shortcomings have simply been accepted. Our data suggest that the relationship between the residual vote rate and the true error rate may be complex. In the aggregate, the inclusion of non-erroneous undervotes may be compensated for by the lack of wrong choice votes, thus, the mean rate for the two may be similar. However, at the individual level this does not appear to hold, as the two measures were only weakly correlated. Strong conclusions based on residual vote rates may be problematic for this reason; more research on this question is clearly warranted.

Overall, usability is an under-valued aspect of the voting process. However, it is a crucial one that has received markedly less study than other concerns such as security, even though security and usability are often related. Our results highlight the need for more empirical study of the factors affecting usability of voting systems, both electronic and otherwise.

Acknowledgements

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. We would also like to thank the VoteBox team for providing the necessary modifications to the VoteBox software.

References

Brace, K. W. (2008). Nation sees drop in use of electronic voting equipment for 2008 election – A first. Manassas, VA: Election Data Services Inc.

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

Byrne, M. D., Greene, K. K., & Everett, S. P. (2007). Usability of voting systems: Baseline data for paper, punch cards, and lever machines. *Human Factors in Computing Systems: Proceedings of CHI 2007*, (pp. 171–180). New York: ACM.

Everett, S. P. (2007). The usability of electronic voting machines and how votes can be changed without detection. Doctoral dissertation, Rice University, Houston, TX.

Everett, S. P., Byrne, M. D., & Greene, K. K. (2006). Measuring the usability of paper ballots: Efficiency, effectiveness, and satisfaction. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, (pp. 2547-2551). Santa Monica, CA: Human Factors and Ergonomics Society.

Everett, S. P., Greene, K. K., Byrne, M. D., Wallach, D. S., Derr, K., Sandler, D., & Torous, T. (2008). Electronic voting machines versus traditional methods: Improved preference, similar performance. *Human Factors in Computing Systems: Proceedings of CHI 2008* (pp. 883-892). New York: ACM.

Goggin, S. N. (2008). Usability of election technologies: Effects of political motivation and instruction use. *The Rice Cultivator: A Student Journal of Public Policy Research*, 1, 30-45.

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

Herrnson, P. S., Niemi, R. G., Hanmer, M. J., Bederson, B. B., Conrad, F. C., & Traugott, M. W. (2008). Voting technology: The not-so-simple act of casting a ballot. Washington, DC: Brookings Institution Press.

ISO 9241-11 (1998). Ergonomic requirements for office work with visual display terminals (VDT) – Part 11: Guidelines on usability. Geneva, Switzerland: International Organization for Standardization.

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.

McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1), 81-90

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

Norden, L., Kimball, D., Quesenbery, W., & Chen, M. (2008). Better Ballots. New York, NY: Brennan Center for Justice at NYU School of Law.

Rubin, A. D. (2006). *Brave new ballot: The battle to safeguard democracy in the age of electronic voting*. New York: Morgan Road.

Sandler, D. R., & Wallach, D. S. (2008). The case for networked remote voting precincts. *Proceedings of the 3rd USENIX/ ACCURATE Electronic Voting Technology Workshop*.

Sandler, D. R., Derr, K., & Wallach, D. S. (2008). VoteBox: A tamper-evident, verifiable electronic voting system. *Proceedings of the 17th USENIX Security Symposium*.

United States Government, 47th Congress. (2002). Help America Vote Act of 2002. Public Law 47-252. Washington, D.C.

Wand, J. N., Shotts, K. W., Sekhon, J. S., Mebane, W. R., Herron, M. C., & 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.