RICE UNIVERSITY

# Usability of New Electronic Voting Systems and Traditional Methods: Comparisons Between Sequential and Direct Access Electronic Voting Interfaces, Paper Ballots, Punch Cards, and Lever Machines

by

**Kristen K. Greene**

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

**Master of Arts**

APPROVED, THESIS COMMITTEE

Michael D. Byrne, Ph. D., Chair
Associate Professor, Psychology

Philip T. Kortum, Ph. D.
Faculty Fellow, Psychology

David M. Lane, Ph. D.
Associate Professor, Psychology

HOUSTON, TEXAS

NOVEMBER 2008

# ABSTRACT

## Usability of New Electronic Voting Systems and Traditional Methods: Comparisons Between Sequential and Direct Access Electronic Voting Interfaces, Paper Ballots, Punch Cards, and Lever Machines

by

Kristen K. Greene

It has been assumed that new Direct-Recording Electronic voting machines (DREs) are superior to the older systems they are replacing, despite a lack of supporting research. The current studies contribute much-needed data on the usability of both older and newer voting systems. Study 1 compared a DRE with a sequential navigation model to paper ballots, punch cards, and lever machines; a DRE with a direct access navigation model was added in Study 2. Changing the navigation style from sequential to direct decreased voter satisfaction and greatly increased undervote errors and intentional abstentions. Premature ballot casting was seen with the direct DRE only. Across both studies, participants were neither faster nor less error-prone with the DREs than the older methods. Nonetheless, they found the sequential DRE significantly more satisfying, an interesting disassociation between preference and performance. Despite voter preferences, the assumption that DREs are superior may be unfounded.

# Acknowledgements

This research would not have been possible without the expertise and assistance of my advisor, Dr. Michael D. Byrne. I am extremely grateful for his careful supervision and continuous support, and simply cannot overemphasize how deeply I appreciate his patient mentorship. I would also like to thank my committee members, Dr. Philip T. Kortum and Dr. David M. Lane, for their invaluable guidance and feedback. Finally, many thanks to my family and friends for their unflagging encouragement.

# TABLE OF CONTENTS

## LIST OF TABLES

# LIST OF FIGURES

# INTRODUCTION

The 2000 presidential election was an extremely high-profile incident in our nation’s voting history, causing people to question the very legitimacy of the election. The infamous Florida butterfly ballot and subsequent recount led to greatly heightened national interest in voting systems and technologies. Much of this interest has been focused on issues of security and audit capacity, with comparatively little interest in usability and human factors issues. This is unfortunate, given that usability is such a critical issue in voting systems.

Even if a voting system existed that was perfectly secure, accurate, reliable, and easily auditable, it would still not be sufficient. The system must also be usable; having a voting system that is secure from tampering means nothing if voters cannot actually use it. Similarly, having voting machines that are reliable and never malfunction will solve nothing if voters cannot use them. Voting systems must be usable by any and every type of voter. A truly usable voting technology should be a walk-up-and-use system, enabling even first-time voters to cast their votes successfully. It should be accessible for those voters with special needs, and it should not disenfranchise particular groups of voters. The Help America Vote Act (HAVA) of 2002 was a landmark piece of legislation in terms of calling public attention to issues such as these, and was a direct consequence of the heightened national interest in electoral reform resulting from the 2000 election.

HAVA secured funding for improvements to election administration (United States Government [US Gov.], 2002), and as a direct result, older systems—such as lever machines and punchcards—are rapidly and widely being replaced with newer electronic voting systems: Direct-Recording Electronic voting machines (DREs). In addition to securing funding, HAVA also established the Election Assistance Commission (EAC). Among its numerous responsibilities, the EAC was charged with the responsibility of developing Voluntary Voting System Guidelines (VVSG), which were passed in 2002 and revised in 2005 (US Gov., 2005). The VVSG covers usability, security, and privacy, while also addressing accessibility for all types of users. These guidelines are a step in the right direction; however, the standards suggested are not mandatory. The decision of whether or not to implement and rigorously follow the VVSG is therefore left up to the discretion of each individual voting district. While this is certainly an issue, an even bigger problem exists: there are very few baseline usability measures for existing voting systems. To determine whether the new electronic voting systems are in compliance with the VVSG and are actually improvements over the older systems, data are crucially needed on the usability of both the older and newer voting methods.

The goal of the two studies reported here is to begin addressing this need for baseline usability data, specifically by comparing the usability of sequential versus direct access navigation models for a prototype electronic voting interface. This research built upon several recent studies that examined the usability of more traditional voting systems, and was one in a series of laboratory studies comparing the usability of multiple voting technologies: paper ballots, mechanical lever machines, punchcard voting systems, and DREs. Usability as it applies to voting systems is unique, in that being objectively usable is not enough; voters must also perceive the system as being usable. The three usability metrics recommended by the National Institute of Standards and Technology

(NIST) address both the objective and subjective usability requirements. NIST recommends using three usability metrics suggested by the International Organization for Standardization (ISO): efficiency, effectiveness, and satisfaction (Laskowski et al., 2004). Efficiency is defined as the relationship between the level of effectiveness achieved and the amount of resources expended, and is usually measured by time on task (Industry Usability Reporting Project, 2001). In other words, efficiency is an objective usability metric that measures whether a user's goal was achieved without expending an inordinate amount of resources. Did voting take an acceptable amount of time? Did voting take a reasonable amount of effort?

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

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

Although previous research in the voting domain has not focused on the three usability metrics of efficiency, effectiveness, and satisfaction, other domains have more widely made use of them. Researchers Frokjaer, Hertzum, and Hornbaek (2000) had subjects solve 20 information retrieval tasks concerning programming problems. They measured efficiency by task completion time and effectiveness by quality of solution, and found that the correlation between efficiency and effectiveness was only negligible in their study. They then went through three years of CHI (Computer Human Interaction) conference proceedings in search of information regarding correlations between the three usability metrics, and concluded "effectiveness, efficiency, and satisfaction should be considered independent aspects of usability and all be included in usability testing." It is very risky to make assumptions about correlations between usability aspects: such correlations depend in a complex way on the application domain, the user's experience, and the use context. Given these considerations, it would seem wise to include all three measures in usability research as it applies to voting systems. As of yet, however, research on the general usability of voting systems is scarce, and research on the more specific aspects of usability is virtually nonexistent.

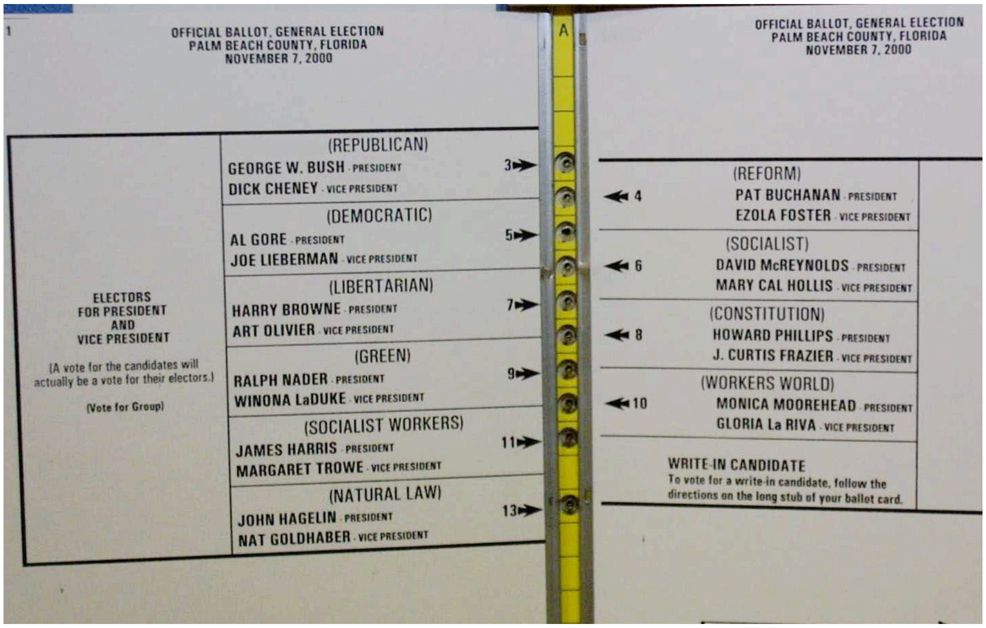
Despite the lack of usability research in the voting domain, numerous other research topics have been studied in great depth, such as the issue of unrecorded votes. The difference between voter turnout and the number of valid votes actually cast for any given race is the typical definition of unrecorded votes, also known as the residual vote (Kimball & Kropf, 2005). The residual vote arises due to a combination of overvoting and undervoting. Overvoting occurs when a voter chooses more than the allowed number of options for a given race; for instance, choosing two candidates rather than one for president would result in neither vote being recorded for that particular office. Similarly, voting both for and against a proposition would result in neither of those votes being recorded. Undervoting occurs when a voter does not indicate a preference for a given race or proposition, i.e. a voter makes no selection. When voters do not record a vote, it is considered an undervote—also known as roll-off. What makes undervoting particularly interesting, yet more difficult to study, is that undervoting can be either intentional or unintentional. Many voters go to the polls planning only to vote for presidential candidates; they intend to leave all subsequent down-ballot races blank, in which case doing so is an intentional omission rather than an error. Yet numerous other voters may plan to vote for every race; any races these voters leave blank would be errors on the part of the voter, perhaps because they accidentally skipped a line, etc. Without a way to assess a voter's intent, there is no way to differentiate between intentional versus unintentional undervoting. The literature offers several possible explanations for the occurrence of both types of undervoting, explanations that are by no means necessarily incompatible with one another.

Bullock and Dunn (1996) have suggested that voter fatigue may be one reason for intentional undervoting. Other research on undervoting has proposed that intentional undervoting may be an indication that a voter was unhappy with the listed candidates, choosing to abstain from voting rather than voting for an undesirable contender (Kimball, Owens, & Keeney, 2004). This idea could be supported by such occurrences as the elevated rate of unrecorded votes (2.8%) found in the states where Ralph Nader was not included on the ballot, versus the lower rate of unrecorded votes (1.7%) found in states which included Nader on their ballots (Kimball, Owens, & Keeney, 2003). These statistics pitted seven states which exclude Nader from their ballots against 24 states which included Nader. What makes the study of undervoting particularly problematic is the decentralized nature of government in America; requirements to report the number of unrecorded votes vary by state, meaning that much research comparing unrecorded votes is by default incomplete.

Another view on undervoting is taken by Wattenberg, McAllister, and Salvanto (2000), who propose that undervoting may be due to voters having inadequate information about candidates. Finally, a third view of undervoting suggests that undervotes may not be intentional, but instead may be due to voter confusion with various technologies. In an analysis of presidential, gubernatorial, and senatorial elections across the United States from 1988 to 2000, Ansolabehere and Stewart (2005) found that punchcards had the highest rates of uncounted votes, followed by DREs, mechanical lever machines, optically scanned ballots, and traditional paper ballots. It is fascinating that the oldest method of voting, by paper ballot, is also the least prone to undervoting. Yet the newest voting technology, DRE, is highly prone to undervoting, second only to punchcards.

Other research on the residual vote (both undervoting and overvoting) has focused on ways in which the rate of unrecorded votes can be mitigated by various ballot design elements. It seems that greater numbers of unrecorded votes occur when candidates for the same office are listed on multiple pages or in multiple columns. (Sinclair & Alvarez, 2004). It is undeniable that all voting systems face challenges with ballot design and presentation. However, only since  the poorly designed butterfly ballot of the 2000 election, have people become so acutely aware that errors in ballot design can actually play a pivotal roll in deciding the outcome of a national election. There is significant statistical evidence to suggest that had precinct-tabulated optical scan ballots been used instead throughout Florida, Al Gore would have been elected as president rather than George Bush, by a margin of over 30,000 votes. (Mebane, 2004). This is arguably due to the fact that such precinct-tabulated systems warn the voter of a spoilt ballot as it is inserted into the scanner. Since the voter is then given an opportunity to correct his/her ballot, these systems essentially offer a usability feature that is standard for many of the interfaces with which humans interact daily: error feedback. Good machine interfaces give the user an effective means of error correction, and most people are undoubtedly familiar with the concept of error feedback and subsequent opportunity for error correction. Error/warning messages on computers, auditory feedback that a card has been left in an ATM, even a reprimand by a supervisor or corrections on paper by a teacher can be considered examples of error feedback. The error feedback offered by precinct-tabulated systems could have mitigated the deleterious effect of the poorly designed butterfly ballot on the results of the 2000 national election.

The exact nature of the butterfly ballot's poor design bears explaining. A single column of punch holes is flanked by two columns of candidate names for the same office (Figure 1).



1

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

ELECTORS
FOR PRESIDENT
AND
VICE PRESIDENT

(A vote for the candidates will actually be a vote for their electors.)

(Vote for Group)

(REPUBLICAN)
GEORGE W. BUSH - PRESIDENT
DICK CHENEY - VICE PRESIDENT
3➡

(DEMOCRATIC)
AL GORE - PRESIDENT
JOE LIEBERMAN - VICE PRESIDENT
5➡

(LIBERTARIAN)
HARRY BROWNE - PRESIDENT
ART OLIVIER - VICE PRESIDENT
7➡

(GREEN)
RALPH NADER - PRESIDENT
WINONA LaDUKE - VICE PRESIDENT
9➡

(SOCIALIST WORKERS)
JAMES HARRIS - PRESIDENT
MARGARET TROWE - VICE PRESIDENT
11➡

(NATURAL LAW)
JOHN HAGELIN - PRESIDENT
NAT GOLDHABER - VICE PRESIDENT
13➡

A

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

⬅ 4
(REFORM)
PAT BUCHANAN - PRESIDENT
EZOLA FOSTER - VICE PRESIDENT

⬅ 6
(SOCIALIST)
DAVID McREYNOLDS - PRESIDENT
MARY CAL HOLLIS - VICE PRESIDENT

⬅ 8
(CONSTITUTION)
HOWARD PHILLIPS - PRESIDENT
J. CURTIS FRAZIER - VICE PRESIDENT

⬅ 10
(WORKERS WORLD)
MONICA MOOREHEAD - PRESIDENT
GLORIA La RIVA - VICE PRESIDENT

WRITE-IN CANDIDATE
To vote for a write-in candidate, follow the directions on the long stub of your ballot card.

*Figure 1.* Butterfly ballot.

These flanking pages are analogous to butterfly wings. This is a non-user friendly design because the punch holes are alternately for the left and right pages of the ballot, which does not map onto the manner in which many voters would vote the ballot. Americans read left to right, top to bottom on a column/page before moving to the adjacent column/page. If voters made their choices in this manner, they would perceive that punching the first hole as corresponding to voting for Bush and the second hole as voting for Gore. Yet in reality, punching the second hole would have resulted in a vote for Buchanan. Although someone voting Republican would still vote correctly for Bush, someone voting Democrat would erroneously vote for Buchanan rather than Gore. This is a large-scale example of a design error that resulted in poor usability for a large population of voters (Wand et al., 2001).

Other graphic design elements, such as the location of instructions, readability, use of shading and bolding, ease of finding the place to mark, and clutter around candidate names, can also influence the residual vote, even widening the racial disparity in unrecorded votes (Kimball & Kropf, 2005). This could play a role in the finding that high presidential overvoting rates were found in Broward and Miami-Dade Florida precincts with greater numbers of blacks and Hispanics (Herron & Sekhon, 2003). The racial gap in unrecorded ballots can depend on the exact type of voting equipment used. For example, research has found this racial disparity to range from four to six percentage points when punchcards or optical scan ballots are used, yet the use of mechanical lever machines and DREs can reduce the racial gap by a factor of ten (Tomz & Houweling, 2003). This difference could be due to the manner in which lever and electronic machines prohibit overvoting, which may have significant implications for election administration. When considering the results from the Tomz and Houweling (2003) study, one must be aware that a very unique dataset was used, comprised only of precinct-level data from South Carolina and Louisiana. These states were chosen because they are the only states which officially report voter turnout by race. While informative, these data may not be representative of the entire United States.

In quite a different area of the U.S., researchers at the University of Maryland (UMD) and the University of Michigan collaborated to conduct an extensive exit poll to assess Maryland’s new voting machines, even including questions related specifically to usability. In the 2002 elections, mechanical lever and punch card voting systems were replaced with electronic touch screen systems, which proved an ideal opportunity to study such things as voter acceptance, voter trust, problems with the new system, and usability issues (Herrnson, Bederson, & Abbe, 2002). It was found that the majority of voters (over 90%) felt the new system was easy to use. Voters also professed greater trust with the new touch screen system than with the older voting systems (91% versus 71%). However, nine percent of voters requested help using the new system. Voters inexperienced with computers found the system more difficult, and those voters who had not attended college were twice as likely to require help as were voters with college educations. Voters aged 65 or older required more help than did voters in other age categories (Herrnson, Bederson, & Abbe, 2002). Herrnson et al. (2002) also found evidence for usability concerns with the new hardware/ software itself. Technical problems were experienced by three out of every one hundred voters. This was primarily because of difficulty activating the voting machine, which required very forceful insertion of a card. Voters also experienced usability problems when navigating between screens, stating that multiple screens were sometimes jumped, and that it was difficult to review their ballots.

In another collaborative UMD study by the Center for American Politics and Citizenship (CAPC) and the Human-Computer Interaction Lab (2006), four vote verification systems and one system without a verification unit were tested for the Maryland State Board of Elections. Verification systems are those electronic voting methods that allow a voter to see a printed record of his or her vote for verification. The logic behind such systems is that voters will be more likely to catch any errors if they are able to check their votes on paper rather than on an electronic screen. The tests conducted for the Maryland State Board of Elections found a perceived trade-off between security and usability, with verification decreasing usability. Many people originally thought that having a DRE with a printer would be a viable solution to the problem of audit capacity in electronic voting systems. However, the Diebold AccuView Printer tested in this study had a paper spool that was not only complex to change, but also required frequent changing (CAPC, 2006). This study also examined a voting system widely touted as a solution to accessibility issues for vision-impaired voters: the MIT audio system, which lets voters hear their choices via a set of headphones. In addition to seeming ideal for blind voters, one might think that forcing voters to hear their choices would enable more errors to be identified. However, in this study, voters expressed significant concerns about privacy, being worried that the volume necessary to hear their choices was loud enough to compromise their privacy. In addition, voters were also concerned about the sanitation of reusing the same headset. The study involved over 800 voters, and due to usability and administrative issues ultimately did not recommend that Maryland invest in any of the vote verification systems tested (CAPC, 2006).

The field studies conducted by CAPC and UMD were landmark steps in gathering large-scale usability data on new electronic voting technology. However, studies like these would still benefit from usability research in more controlled laboratory settings. Laboratory usability studies would be an extremely informative complement to field research, with the potential to offer converging evidence by providing more thorough data, as well as the ability to manipulate variables of interest. One might propose that the ultimate aim of usability research in the voting arena is actually two-fold: 1) to empirically assess the comparative usability across the various voting methods and 2) using such assessments of usability, make design recommendations for improvements to existing systems, as well as for better development of new systems.

In pursuit of the first of these two research goals, baseline usability data for older voting systems is obviously necessary. Only when such data is gathered for these older voting methods, such as lever machines, paper ballots, and punchcards, may justified conclusions be drawn regarding whether the newer, electronic voting systems are indeed improvements over more traditional methods. Recent laboratory studies have successfully begun addressing this need for baseline usability data. Everett, Byrne, & Greene (2006) used measures of efficiency, effectiveness, and satisfaction to compare the usability of three different types of paper ballots. The three types of paper ballots evaluated by were open response ballots, bubble ballots, and arrow ballots. Every participant voted on all three ballot types. Each participant saw either realistic candidate names or fictional candidate names, but never both. Realistic names were those of people who had either announced their intention to run for the office in question, or who seemed most likely to run based on current politics. Fictional names were produced by a random name generator. Participants were also provided with one of three different types of information. They were given a slate that told them exactly who to vote for, given a voter guide, or were not given any information about candidates at all.

The three dependent measures used by Everett, Byrne, & Greene (2006) were ballot completion time, error rates, and satisfaction. These addressed the NIST's suggestions for the usability metrics of effectiveness, efficiency, and satisfaction. Ballot completion time measured how long it took the participant to complete his or her vote. No significant differences were found between ballot types (bubble, arrow, or open response) or candidate types (realistic or fictional) on the ballot completion time measure. Significant differences were observed between information types; participants given a voter guide took reliably more time to complete their ballots than those who received a slate or those who did not receive any information about candidates. The second dependent measure, errors, were recorded when a participant marked an incorrect response in the slate condition or when there was a discrepancy between their responses on the three ballot types. Error rates can be considered in two ways: by race, and by ballot. "By race" error rates were low, with participants making errors on less than 4% of the races, a race being considered either a candidate race or a proposition. When considering "by ballot" error rates, one classifies each ballot as either error free, or containing at least one error. These error rates were surprising, in that over 11% of all ballots contained at least one error. The reason this is especially disturbing is that participants in this study were Rice University undergraduates, an extremely intelligent, technologically experienced, and well-educated sample, the majority of whom come from families of high socioeconomic status. Finally, the third dependent measure, satisfaction, was assessed by the participant's responses to the SUS. The bubble ballot produced significantly higher SUS scores than the other two ballots, while the open response and arrow ballot were not significantly different from one another. The authors concluded that all three ballot types were acceptable in terms of efficiency and effectiveness, while voters were much more satisfied with the bubble ballot.

In a subsequent study, Greene, Byrne, and Everett (2006) used the same three usability measures, efficiency, effectiveness, and satisfaction, to assess the usability of mechanical lever machines, in addition to measuring usability for two of the three previously studied paper ballots. Only fictional names were used in this study, since Everett, Byrne, & Greene (2006) found no significant differences on ballot completion times, error rates, or satisfaction ratings between realistic or fictional candidate names. This decision was made to prevent advertising campaigns and media coverage from differentially affecting data from one study to the next. In addition, the continued use of realistic candidate names would have quickly rendered the study materials obsolete, making comparisons between studies with new materials difficult. Using fictitious names circumvents both these issues for the current study and future research. The bubble and arrow ballots received the highest and lowest satisfaction ratings, respectively, and no significant differences were found between ballot completion times and error rates for the three paper ballot types used in the prior study. Therefore, only the bubble and arrow ballots were used in the second study. With the addition of another voting technology, mechanical lever machines, it was necessary to choose only two of the three previously studied paper ballots in order to keep the total number of voting methods at three. This is important because it allows the voter’s intent to be determined by a clear majority, i.e. if a voter chose the same candidate with two of the three voting methods, one can be confident that the discrepant vote is truly an error.

Greene, Byrne, and Everett (2006) again found no reliable differences in efficiency between voting methods. Similarly, no reliable differences in effectiveness between voting methods were found. This was the case both when considering “by race” error rates and “by ballot” error rates. Despite this, the by ballot error rate was much higher than that found in their prior study (Everett, Byrne, & Greene, 2006), jumping from 11% to nearly 16% of ballots containing at least one error. This is presumably due the fact that they sampled from the general Houston population rather than using participants solely from Rice University. Reliable differences in satisfaction were again found; people were most satisfied with the bubble ballot and least satisfied with the lever machine. The persistence of greater satisfaction with the bubble ballot across these two studies is interesting, suggesting that their initial finding of higher satisfaction with the bubble ballot was not merely due to their use of Rice undergraduates as participants.

Wishing to expand upon their prior research while still allowing comparisons across studies, Byrne, Greene, and Everett (2006) utilized the methodology previously described, with one notable exception: the addition of a punch card voting system as a fourth within-subjects condition. Along with the addition of punch cards, the other three within-subjects conditions were the arrow ballot, bubble ballot, and lever machine, which were identical to those used in prior studies. Two information conditions, voter guide and slate, again served as the between-subjects experimental manipulation. All materials—voter guides, fictional candidate names, and propositions—were identical to ones from their two previous studies. Again following the methodology from their earlier studies, efficiency, effectiveness, and satisfaction were dependent variables used to assess usability across the four voting methods.

Just as in the two studies previously described, Byrne, Greene, and Everett (2006) found that voters took about the same amount of time to complete a ballot regardless of the voting method used. Yet with this study, a reliable linear effect of education was found on ballot completion times, with lower levels of education associated with longer completion times. However, a reliable effect of education was not found for error rates, making it unclear exactly what underlying facet of education was actually responsible for the varying completion times. "Per-race" error rates between the voting methods were reliably different, as were error rates between information conditions, with fewer errors made in the slate condition than in the guide condition. "By ballot" error rates, which considered each ballot as either error-free or as containing at least one error, were not reliably different between voting methods. Although 26% of all ballots finished (78 of 300) contained at least one error, no voting method was more or less likely to generate ballots with at least one error. The error rates seen in this study were almost 4% greater than the 16% and 11% found in prior research (Greene, Byrne, and Everett, 2006; Everett, Byrne, & Greene, 2006). This was likely due to the use of participants more representative of the general voting public, as they were sampled from the general Houston population. Even with this extremely diverse sample, the bubble ballot again scored the highest in subjective usability ratings, a finding which has remained consistent across the series of three voting system usability studies.

The three previously-described studies were successful in gathering baseline usability data for paper ballots, lever machines, and punch cards (Everett, Byrne, & Greene, 2006; Greene, Bryne, and Everett, 2006; Byrne, Greene, & Everett, 2006), a necessary first step in being able to compare the usability of more traditional voting systems to newer electronic ones. Obviously, the next logical step in achieving this end comparison would be to collect data on the usability of DREs. However, most DREs currently on the market would not guarantee access to the data we would want, so it was necessary to build our own electronic voting system. There are important benefits to designing and building an electronic voting interface from scratch. With complete control over the design process, one has the ability to manipulate features of interest and record detailed data on the effects of such manipulations.

The overarching research interest guiding these studies was a desire for an empirical comparison of sequential versus direct access navigation models, specifically as they pertain to electronic voting systems. There were two studies; the first compared a DRE with a sequential navigation model to paper ballots, punch cards, and lever machines. In the second study, a new type of DRE was added, one with a direct access navigation model.

# STUDY 1

## *Method*

### *Participants*

The participants for Study 1 were all Rice University undergraduate students: 25 male and 25 female. Ages ranged from 18-27 years, with a mean age of 19.46 (*SD*=1.66

years). All participants had normal or corrected-to-normal vision and were fluent English speakers. The majority of students reported spending between ten and 40 hours a week on computers (Table 1), and 45 of 50 reported that they could touch-type. As would be expected of young college students, many rated themselves high in computer expertise. The mean self-rating of computer expertise was 6.58 (*SD*=1.47) on a 4-point Likert scale, with one representing "novice" and 4 representing "expert."

*Table 1*. Hours per week participants spend on computers.

| Hours | Number of people | Percentage of people |
|---|---|---|
| Less than 5 | 1 | 2% |
| 5 to 4 | 6 | 12% |
| 4 to 20 | 16 | 32% |
| 20 to 30 | 4 | 20% |
| 30 to 40 | 14 | 28% |
| Over 40 | 3 | 6% |

On average, participants had voted in only .56 national elections, ranging from zero to three, and had voted in an average of 2.27 non-national elections, ranging from zero to 4.

*Design*

A mixed design with one within-subjects variable and two between-subjects variables was used. The within-subjects variable was called "subjects' three ballots," as each participant completed the same 27-race ballot three times: first with a DRE, second with a non-DRE voting method, and third with another DRE. There were two slightly different versions of the same DRE; one DRE utilized a sequential navigation model with state information, while the other DRE did not provide any state information. Counterbalancing was used to determine whether the sequential DRE with or without state information was used first versus third. Three ballots were used in order to determine a voter’s intent by majority rule, which was necessary to evaluate errors.

In addition to voting twice with the DRE (one time with state information, another time without it), participants also used one of three other non-DRE methods: paper ballots, punch cards, or lever machines. The between-subjects manipulation was that one third of all participants voted on a mechanical lever machine as their intervening non-DRE technology, a different third of all participants voted on a punchcard, and the remaining third of all participants on a bubble ballot. The assignment of participants to each level of this between-subjects variable was random.

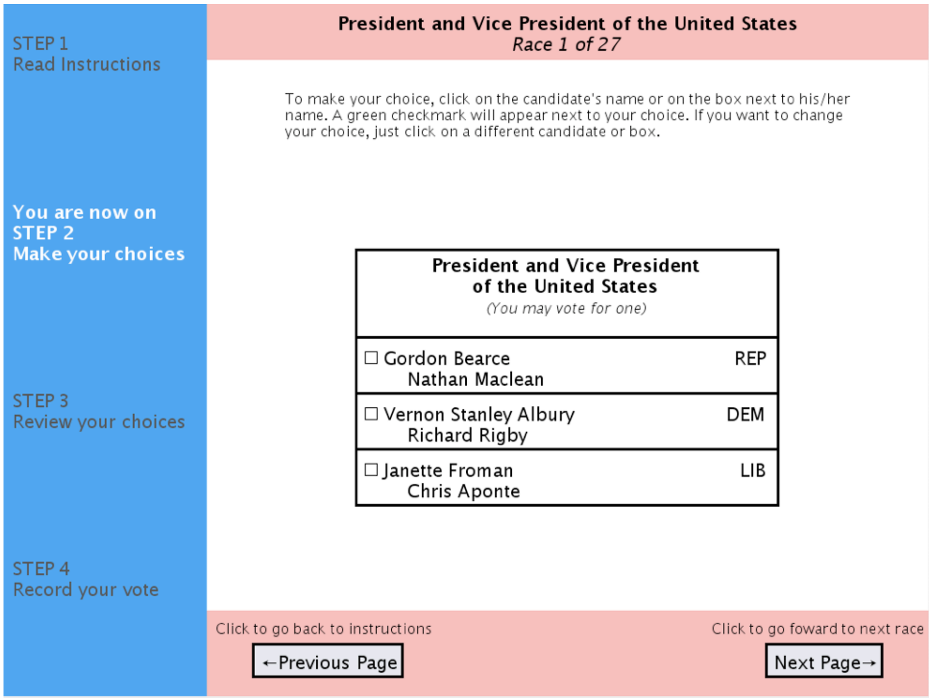
The second between-subjects manipulation was information condition. Each participant was in only one of three possible information conditions: the undirected condition, the directed with no roll-off condition, and the directed with moderate roll-off condition. The assignment of participants to each information condition was random. The voter guides were identical to ones used in the previously described studies; those voter guides were based on guides produced by the League of Women Voters. It was left completely up to the participant whether she/he chose to actually read and use the voter guide. In both of the directed information conditions, participants were given a sheet of paper that listed exactly who the participant was to vote for, as well as whether to vote yes/no on the propositions. Participants kept this paper with them the entire time they were voting. In the directed with no roll-off information condition, participants were instructed to vote in each of the 27 races on a ballot, whereas in the directed with moderate roll-off condition, participants were instructed to intentionally omit certain races. This more closely mimicked the real-world voting patterns in which people do not vote for every race on the ballot. To determine which races were omitted for the directed with moderate roll-off condition, a random number generator was used.

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 ballots. Satisfaction was measured by the participant's responses to the SUS; participants filled out three separate SUS questionnaires, one for each ballot they completed.

### *Materials and Procedure*

The same punch cards, mechanical lever machine, and bubble ballots used in prior studies were used in this research. Each voting method used the same races, fictional candidate names, and propositions used by Everett, Byrne, & Greene (2006). Candidate names were produced by a random name generator (http://www.kleimo.com/random/name.cfm). The 21 offices ranged from national races to county races, and the six propositions were real propositions that had previously been voted on in other counties/states, and that could easily be realistic issues for future elections in Harris County, TX, the county in which these two studies were conducted.

The DREs used were versions of software called VoteBox. Both versions of VoteBox ran on the same hardware, a desktop PC running Windows as its basic operating system. One version provided state information to the user, while the other did not. The state information was clearly visible underneath the title of a race at the top of the screen, and was provided in a K of N format. Specifically, the state information said “Race K of 27,” as there were a total of 27 contests on the ballot (Figure 2).



**President and Vice President of the United States**
*Race 1 of 27*

To make your choice, click on the candidate's name or on the box next to his/her name. A green checkmark will appear next to your choice. If you want to change your choice, just click on a different candidate or box.

**President and Vice President of the United States**
*(You may vote for one)*

| | |
|---|---|
| □ Gordon Bearce<br>Nathan Maclean | REP |
| □ Vernon Stanley Albury<br>Richard Rigby | DEM |
| □ Janette Froman<br>Chris Aponte | LIB |

Click to go back to instructions

Click to go foward to next race

*Figure 2.* Sequential VoteBox presidential race screen, with state information.

Participants first went through the informed consent procedure with the experimenter, and then read the instructions on their own. This was followed by an opportunity for them to ask any questions they had, and then the experimenter reviewed with them the most important parts of the instructions. It is important to note that participants in the undirected versus directed information conditions received slightly different instructions. Instructions for the undirected condition strongly emphasized that participants were to be consistent across voting methods. In other words, people were explicitly instructed to vote for the same candidates on each of the three ballots; these instructions were repeated each time a participant received one of their three ballots. In the directed information conditions, participants were given a sheet of paper that listed the choice they should make for each race on a ballot. Instructions for the directed information conditions emphasized that participants were to make exactly those choices indicated on the piece of paper, and were to keep that paper with them and use it while voting on each of their three ballots.

Following the informed consent procedure and discussion of instructions, those participants in the directed information conditions were given their list of candidates and then began voting immediately on their first DRE. Those participants in the undirected information condition were instead given ample opportunity to read over the voter guide (if they chose to do so) before beginning to vote. Participants were only given one ballot, i.e. shown to one voting station, at a time. The DRE software recorded ballot completion times automatically, but for the bubble ballot, punchcard, and lever machine, the experimenter started a stopwatch as soon as a participant started his/her intervening ballot, and stopped the watch as soon as a participant called out that s/he was done. Participants stood throughout the entire voting process. The VoteBox electronic voting interfaces, paper ballots, and punch cards each had their own separate voting station, a waist-high table with the appropriate technology sitting atop. The lever machine was a stand-alone unit, so provided its own voting station, since participants stood behind its curtains and reach up to pull the levers downwards.

Directly after each of their three ballots, participants filled out a new SUS that evaluated the voting method they had just used. The SUS is a scale that has been tested and verified in multiple domains. It is a Likert scale, which has people answer questions using a scale of one to five, ranging from strongly disagree to strongly agree. The SUS includes questions such as “I thought the system was easy to use” and “I felt very confident using the system.” Answers to ten such questions are then combined to yield a single number; the higher the final number, the higher the degree of satisfaction a person experienced (Brooke, 1996).

After completing all three ballots, participants then filled out a survey packet, which included both general demographic questions and questions more specific to voting itself. Voting specific questions asked about people’s previous voting experience, as well as their opinions on the exact voting methods that they used in the study.

## *Results, Study 1*

### *State information*

Only 16 of 50 participants (32%) reported noticing the state information. Of the 16 who noticed the state information, 12 (75%) reported preferring the VoteBox version with state information to the one without. Eleven of 16 (69%) people reported that they liked knowing how many races remained on the ballot.

*Usability results*

*Efficiency*

Data from three 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. For all the following statistical tests reported, the sequential Bonferoni method was used to correct for multiple comparisons when applicable.

As is common with human subjects performing timed tasks, there was a slight positive skew to the distribution of ballot completion times (Figure 3).

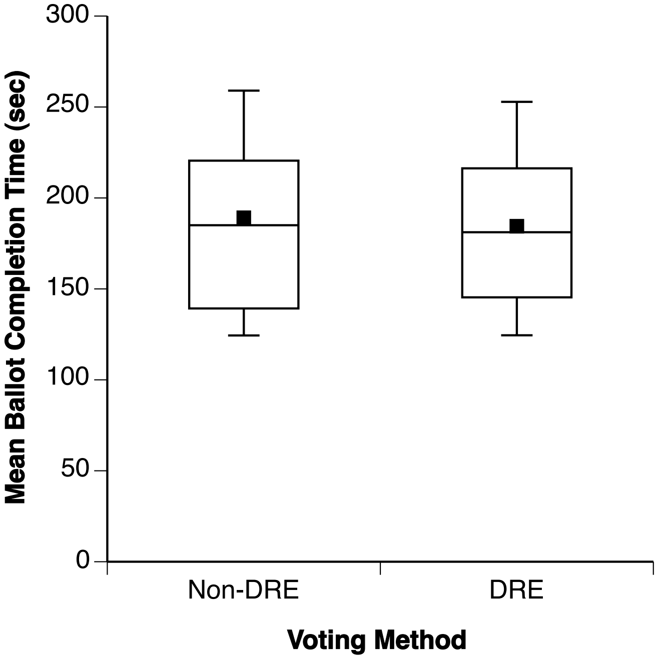


*Figure 3.* Distributions of ballot completion times for DRE and non-DRE voting methods.

There were no noticeable differences in ballot completion times between DREs and paper ballots, punch cards, or lever machines (Figure 4).

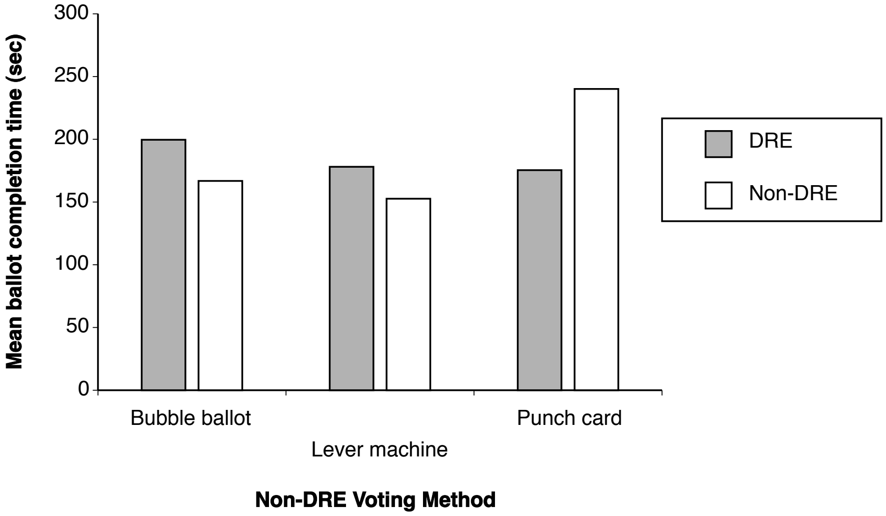


*Figure 4.* Mean ballot completion times for each DRE/non-DRE pair.

When examining times for each of the three information conditions, there were still no significant differences in times between voting methods. In other words, ballot completion times were fairly stable regardless of whether people were in the undirected or directed information conditions.

As can be seen in Figure 5, punch card times were much longer than DRE times (about 240 versus 184 seconds respectively), whereas times for the bubble ballot and lever machine did not differ greatly from DRE times. A 3 (information type) x 3 (non-DRE voting method) x 3 (subjects' three ballots) ANOVA found this interaction between non-DRE voting method and subjects' three ballots to be reliable, *F*(4, 76) = 2.99, *p* = . 02. More specifically, only when the non-DRE method used was the punch card, was there a down-up-down (corresponding to DRE-punch card-DRE) pattern in ballot completion times. The post-hoc interaction contrast testing this effect was statistically reliable, $F(1, 38) = 25.35, p < .001$.

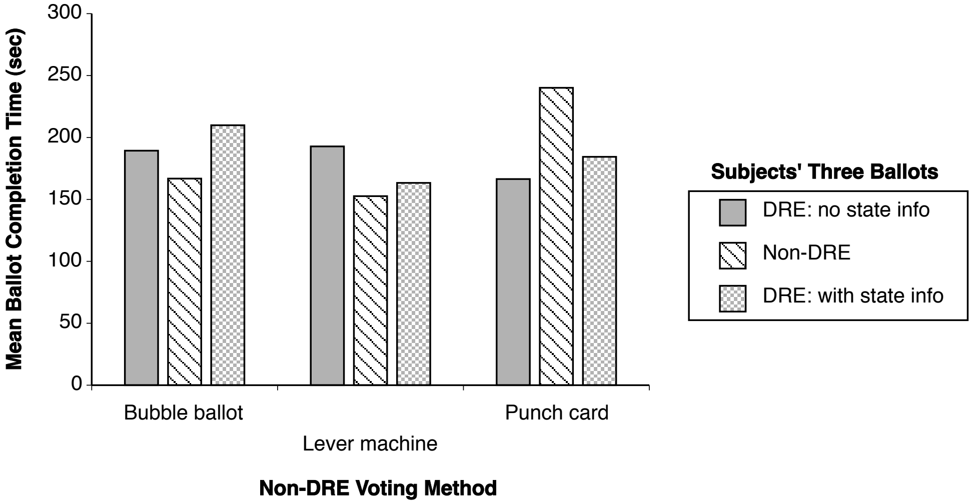


*Figure 5.* Interaction between subjects' three ballots and non-DRE voting method on ballot completion times.

*Effectiveness*

Data from seven participants were not included in the effectiveness analyses, six of whose data were discarded due to program errors in data recording, and one of whose data were discarded due to failure to comply with instructions.

One can conceptualize error rates in two different ways: by race and by ballot. When considering errors by race, the error rate was determined by dividing the number of errors made by the total number of opportunities for error. Since the ballot used had 27 races on it, and each participant voted a total of three different times, every person was confronted with 81 total opportunities for error. Error rates were extremely low overall, less than 1%, and were comparable between all voting methods (Figure 6).

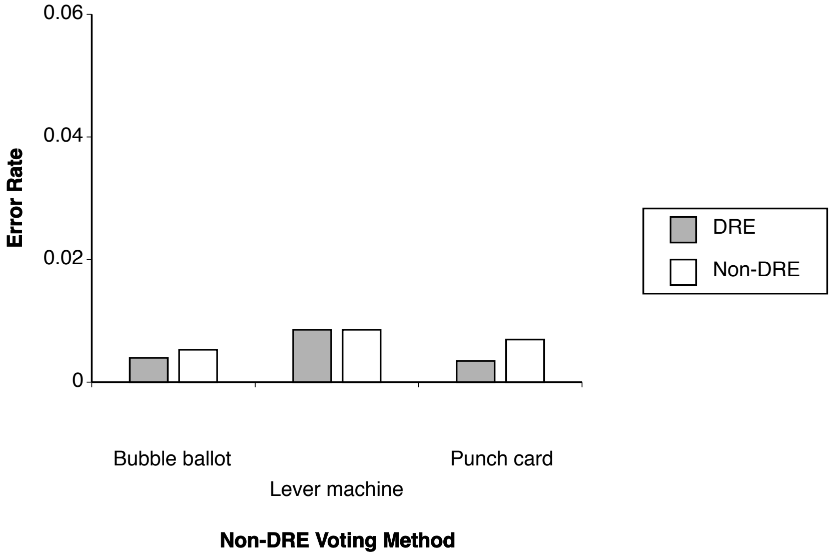


*Figure 6.* Error rates for each pair of DRE/non-DRE voting methods.

Error rates were not influenced by information condition; whether subjects were allowed to make their own choices versus instructed who to vote for had no effect on their propensity to err (Figure 7). This was true both across and within information conditions.

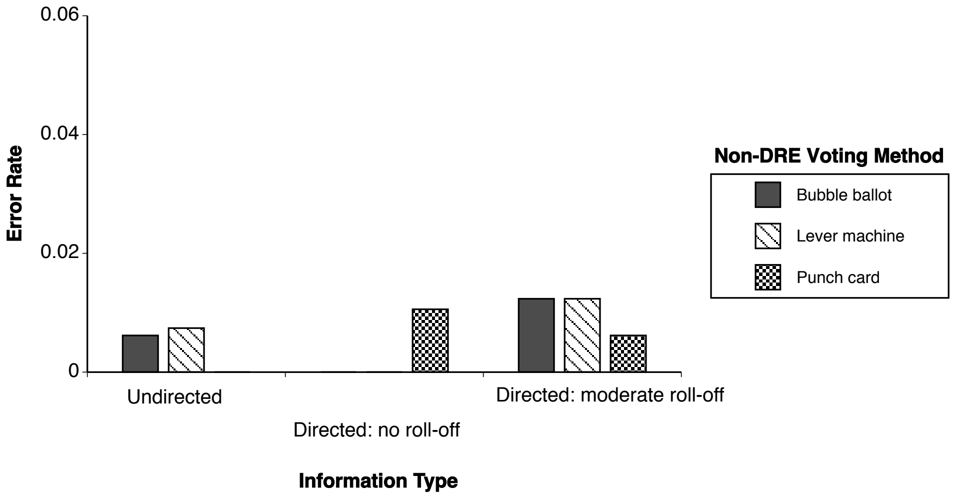


*Figure 7.* Error rates by information type and non-DRE voting method.

As would be predicted from examining Figures 6 and 7, a 3 (non-DRE voting method) x 3 (information type) x 3 (subjects' three ballots) ANOVA on error rates did not find any significant main effects or interactions.

Another way to examine errors is by ballot; each ballot either does or does not contain at least one error. Seventeen out of 127 ballots (13%) contained at least one error. Ballot error frequencies are presented in Table 2.

*Table 2.* Frequency of ballots containing at least one error.

| | Errors | | |
|---|---|---|---|
| | None | At least 1 | Total |
| DRE | 74 | 4 | 84 |
| Bubble | 12 | 2 | 14 |
| Lever | 12 | 4 | 16 |
| Punch card | 12 | 1 | 13 |
| Total | 14 | 17 | 127 |

*Satisfaction*

Data from all participants were used in the following satisfaction analyses. When a person's non-DRE voting method was the lever machine or punch card, they showed a pronounced difference in satisfaction between either of those two technologies and the DRE, preferring the DRE. However, when participants used a bubble ballot, they did not exhibit such a strong preference for the DRE (Figure 8).

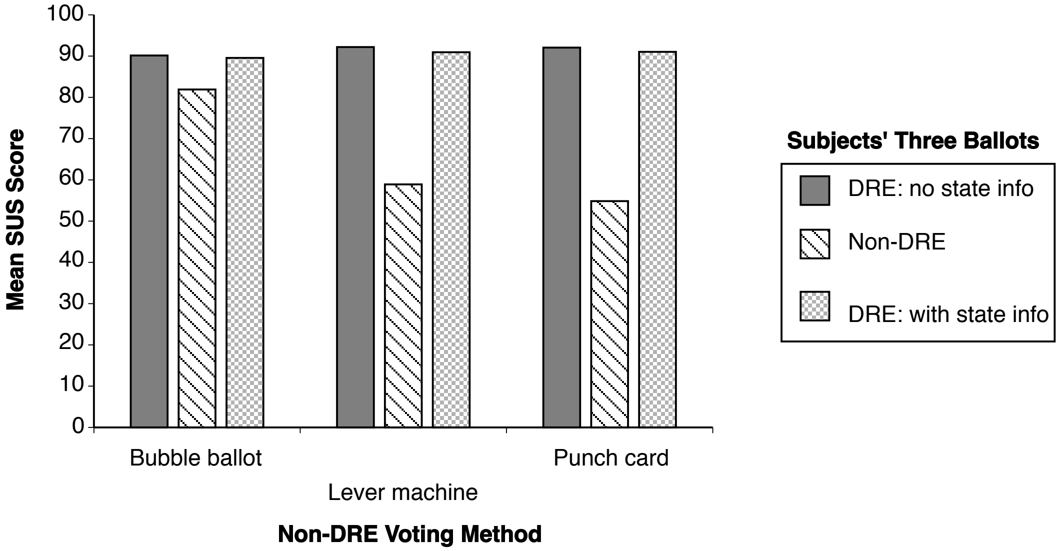


*Figure 8.* Interaction between subjects' three ballots and non-DRE voting method on SUS scores.

The interaction (shown in Figure 8) between non-DRE voting method and subjects' three ballots was reliable, *F*(2.46, 50.37) = 15.21, *p* < .001, as found by a 3 (non-DRE voting method) x 3 (information type) x 3 (subjects' three ballots) ANOVA on SUS scores. The Greenhouse-Geiser method was used to correct for severe sphericity violations. Simple main effects tests showed that the interaction between non-DRE voting method and subjects' three ballots was responsible for a reliable main effect of subjects' three ballots on SUS scores, where SUS scores for the non-DRE voting methods overall were lower than those for both the DREs, *F*(1.23, 50.37) = 116.73, *p* < .001 (Figure 9).

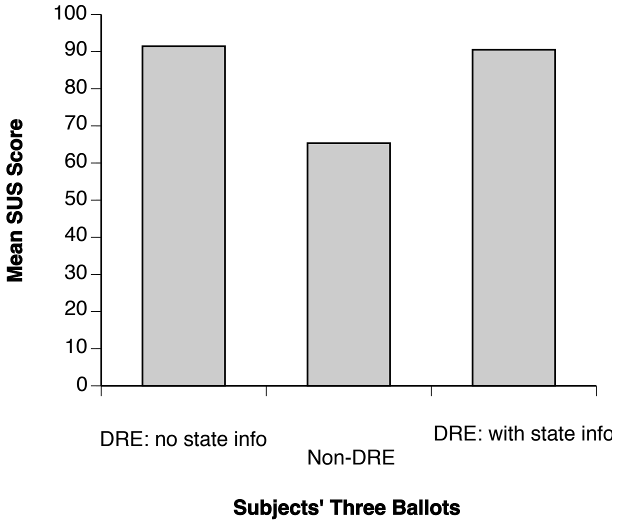


*Figure 9.* Main effect of subjects' three ballots on SUS scores.

Overall, bubble ballot SUS scores were higher than those for the lever machine or punch card (Figure 10). This resulted in a reliable main effect of non-DRE voting method on SUS scores, *F*(2, 41) = 4.04, *p* = .03. More explicitly, the mean bubble ballot SUS score was greater than the mean of the lever machine and punch card scores, as shown by a reliable posthoc contrast, *F*(1, 41) = 28.98, *p* < .001.

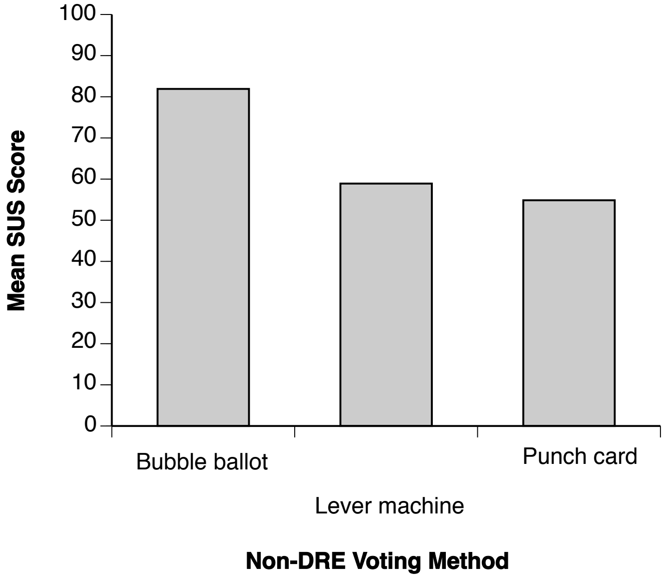


*Figure 10.* Main effect of non-DRE voting method on SUS scores.

*Discussion, Study 1*

The presence or absence of state information had no reliable effects on efficiency, effectiveness, or satisfaction. This may have been due in part to its location at the top of the screen. Although this location was chosen in order to display the state information near the instructions, it may have inadvertently resulted in "banner blindness" (Benway & Lane, 1998). This phenomenon is a pervasive problem with websites, where users commonly fail to notice information when it is presented as a header at the top of the page. The fact that only 16 of 50 participants (32%) reported noticing the state information would suggest banner blindness was likely an issue. Although 12 of those 16 participants reported preferring the DRE with state information over the DRE without it, this preference was not actually reflected in their SUS scores.

However, SUS scores were significantly affected by the type of voting method used. Of the three non-DRE methods, bubble ballots received significantly higher satisfaction ratings than both lever machines and punch cards, which was consistent with prior findings (Byrne, Greene, & Everett, 2007; Everett, Byrne, & Greene, 2006; Everett et al. 2008; Greene, Byrne, & Everett, 2006). Differences in SUS scores within DRE/non-DRE pairs were mitigated by the type of non-DRE used. A significant preference for the DRE was seen only when it was paired with lever machines or punch cards; no such preference for the DRE occurred when it was paired with bubble ballots.

When comparing efficiency between voting methods, no significant differences were seen between DREs and non-DREs overall. The only reliable differences in ballot completion times within specific DRE/non-DRE pairs were found when the non-DRE method used was the punch card, with the punch card times being the longer of the pair. This effect has not been replicated; no reliable timing differences between voting methods have been seen in prior research. However, the lack of significant timing differences between the various information conditions was indeed consistent with previous results. As in prior research, effectiveness was not significantly influenced by information condition, nor did it vary as a result of the non-DRE method used (Byrne, Greene, & Everett, 2007; Everett, Byrne, & Greene, 2006; Everett et al. 2008; Greene, Byrne, & Everett, 2006). The low overall error rates in Study 1 were also quite comparable to those reported previously with Rice undergraduate participants (Everett, Byrne, & Greene, 2006; Greene, Byrne, & Everett, 2006), as was the percentage of ballots containing at least one error.

While informative, the data from Study 1 were not necessarily representative of the true voting public, as all participants in the study were Rice University undergraduate students. The need for a more heterogeneous and representative sample was successfully addressed in Study 2.

## STUDY 2

### *Method*

#### *Participants*

The participants for Study 2 were all recruited from the general Houston population via newspaper advertisements. Participants were paid $25 for their participation in the one-hour study. Of the 64 participants, 30 were male, and 34 female. Ages ranged from 18 to 77 years, with a mean age of 50.3 (*SD* = 14.8). Forty-four participants stated they could touch-type, and the mean self-rating of computer expertise was 5.65 (*SD* = 2.72) on a 4 point Likert scale, with one representing "novice" and 4 representing "expert." Participants spent a fair amount of time on computers each week (Table 3).

*Table 3.* Hours per week participants spend on computers.

| Hours | Number of people | Percentage of people |
|---|---|---|
| Less than 5 | 14 | 22% |
| 5 to 4 | 14 | 22% |
| 4 to 20 | 12 | 19% |
| 20 to 30 | 4 | 16% |
| 30 to 40 | 3 | 5% |
| Over 40 | 7 | 11% |
| Non-response | 4 | 6% |

On average, participants had voted in 9.34 (*SD* = 8.59) national elections, ranging from zero to 30, and had voted in an average of 9.72 (*SD* = 11.52) non-national elections, ranging from zero to 50. As would be expected when using a sample more representative of the general voting public, there was a fair amount of variability in ethnicity, annual income, and education. All subjects were US citizens fluent in English.

### *Design*

In Study 2, a new information condition was added, called "directed with additional roll-off." Participants in that condition were instructed to abstain from voting in an even larger number of races than the number skipped by participants in the directed with moderate roll-off condition. In Study 2, a fourth of participants were in the new directed with additional roll-off information condition, a fourth in the directed with moderate roll-off condition, a fourth in the directed with no roll-off condition, and the remaining fourth in the undirected condition. Assignment to information conditions was random.

The most important difference between Studies 1 and 2 was the addition of a new DRE, one with a direct access navigation style. Therefore navigation type was a between-subjects variable in Study 2; half of all participants used a sequential DRE, while the remaining participants used a direct DRE. The same two versions of the sequential DRE from Study 1 were used, one with state information and the other without. There were also two versions of the direct DRE (one with state information and the other without).

### *Materials and Procedure*

With the exception of the new direct access DREs, the materials used in Study 2 were identical to those in Study 1. State information on the direct DRE was presented in the form of "Voted in K of 27," and was located at the top of the "Main Page" screen just underneath the instructions (Figure 11). The Main Page looked almost identical to the review screen used in Study 1. The direct DRE was somewhat analogous to a webpage, in that all race titles appeared on its Main Page and acted as hyper-links. This meant that by clicking on the race titles, a user could pick and choose exactly which races s/he wished to see, with the option to jump straight to the Record Vote screen at any time. This was in sharp contrast to the sequential DRE, which forced users to page sequentially through every race on the ballot before they could cast their votes.

## *Results, Study 2*

### *Usability*

### *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. For all the following statistical tests reported, the sequential Bonferoni method was used to correct for multiple comparisons when applicable.

The distributions of ballot completion times for Study 2 (Figure 11) were more positively skewed than those of Study 1 (Figure 3).

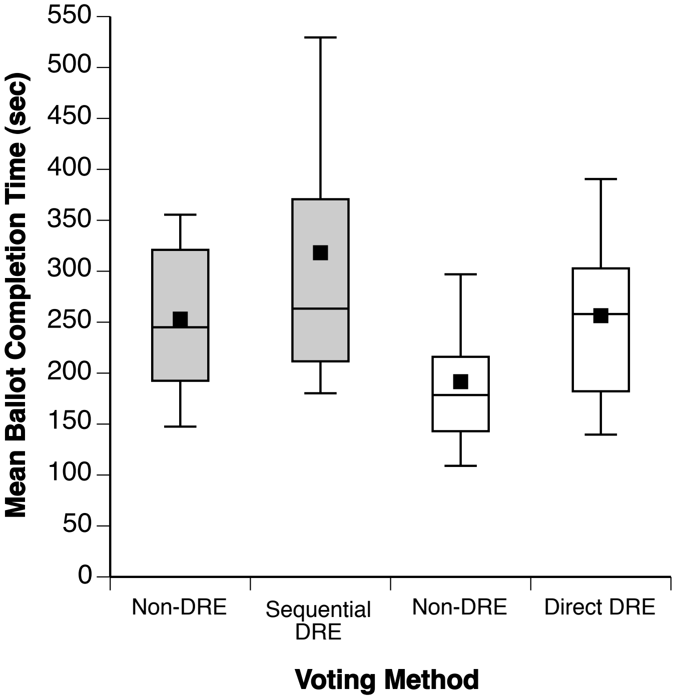


*Figure 11.* Distribution of ballot completion times for DRE/non-DRE voting method pairs.

In the undirected information condition, ballot completion times for the sequential DRE were over twice as long as times for the direct DRE: 453 seconds (*SD*=123) versus only 205 (*SD*=119) respectively (Figure 12).

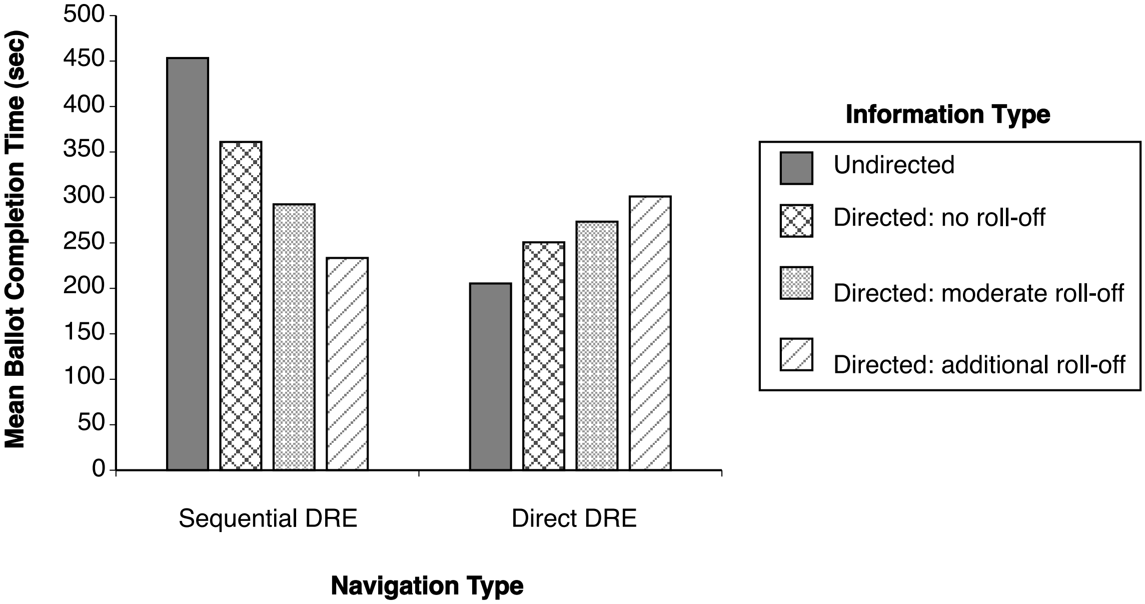


*Figure 12.* Interaction between navigation type and information condition on ballot completion times.

A 4 (information type) x 3 (non-DRE voting method) x 2 (navigation type) x 3 (subjects' three ballots) ANOVA found the interaction between navigation type and information condition to be statistically reliable, *F*(3, 34) = 5.26, *p* = .004. Simple main effects tests showed that only in the undirected information condition was there a significant effect of navigation type on ballot completion times.

Overall, participants using the direct DRE were faster than those who used the sequential DRE (Figure 13), attributable to the extremely long sequential DRE times in the undirected information condition just described. The main effect of navigation type on ballot completion times was statistically reliable, *F*(1, 34) = 11.4, *p* = .002.

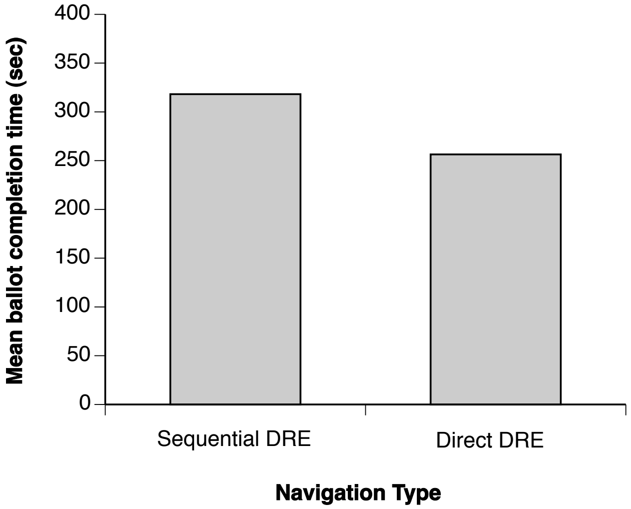


*Figure 13.* Main effect of navigation type on completion times.

The long sequential DRE times in the undirected information condition also meant that in general, participants voted more quickly with the non-DRE methods than with the DREs (Figure 14), resulting in a reliable main effect of subjects' three ballots on completion times, F(2, 68) = 5.57, p = .006. More explicitly, because participants voted first with a DRE, second with a non-DRE, and third again with a DRE, there was a significant up-down-up pattern in the ballot completion time data overall, t(57) = 5.02, p < .001 (tested by a post-hoc contrast).

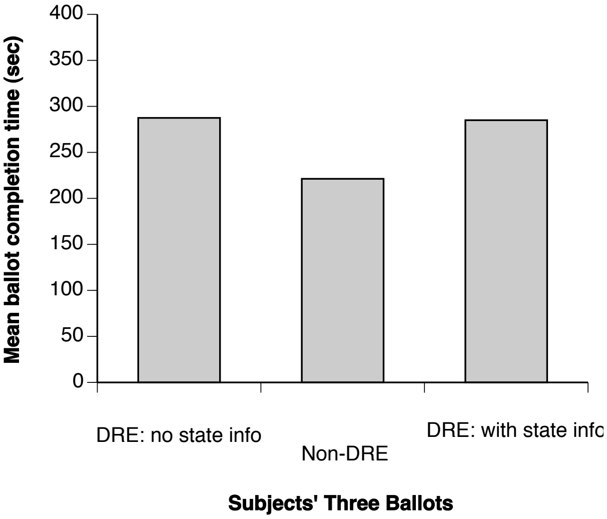


*Figure 14.* Main effect of subjects' three ballots on completion times.

Although a large difference in ballot completion times between DREs was found in the undirected information condition, no significant differences in times between non-DRE voting methods were seen in any of the four information conditions (Figure 15).

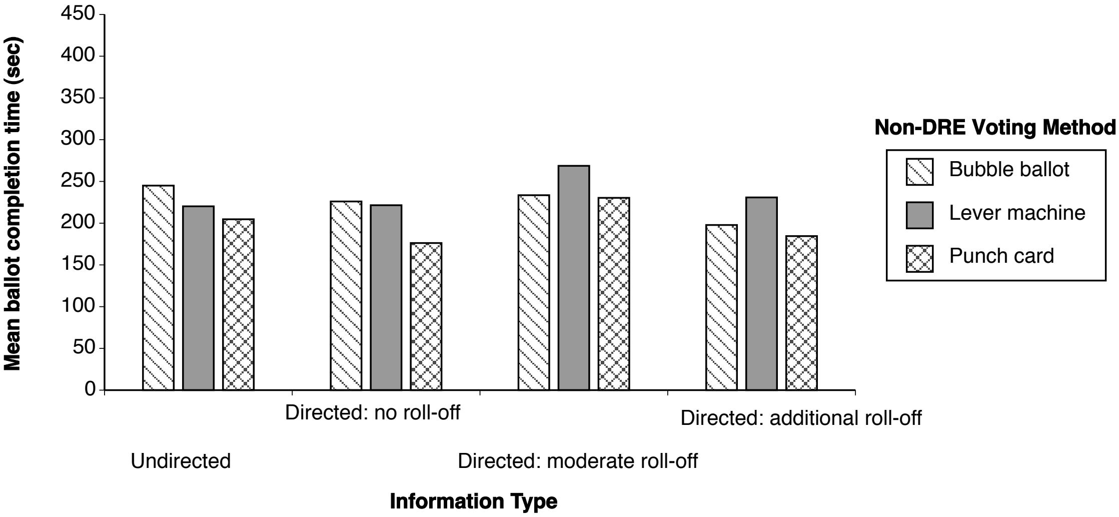


*Figure 15.* Ballot completion times by information condition and non-DRE voting methods.

Furthermore, there were no significant differences in ballot completion times between the sequential DREs and paper ballots, punch cards, or lever machines, nor were there differences between direct DREs and the non-DRE voting methods (Figures 16 and 17).

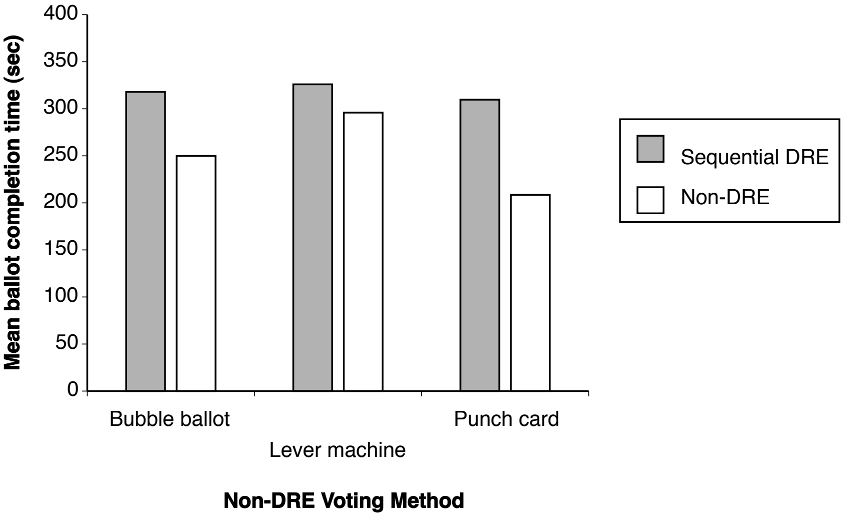


*Figure 16.* Mean ballot completion times for each pair of sequential DRE/non-DRE voting methods.

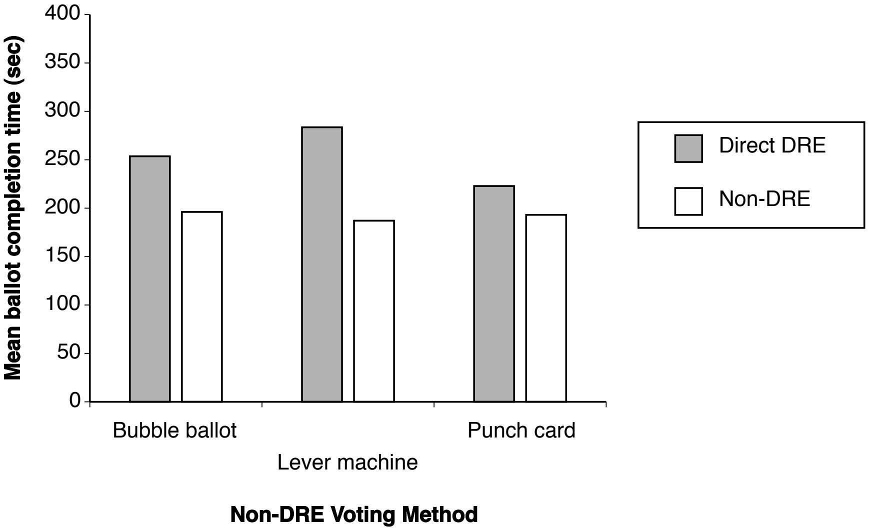


*Figure 17.* Mean ballot completion times for each pair of direct DRE/non-DRE voting methods.

*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 voting methods they used. This definition is in keeping with previous work (Byrne, Greene, & Everett, 2007). This study examined extra vote errors, undervote errors, and wrong choice errors separately. If a voter chose a candidate for a race s/he had planned to omit, this was considered an *extra vote* error. In this study, the intention to omit a race could have been due to instructions given in the directed information conditions or due to a participant's own preference in the undirected condition. An *undervote* error occurred when a participant failed to make a selection for a race in which they had intended to vote. In contrast with an undervote error, an *intentional undervote* occurred when a voter correctly abstained from voting in a race s/he had intended to skip. A *wrong choice* error occurred if a participant chose a candidate other than the one s/he originally intended to choose. The three aforementioned error types have also been combined into one "total error" error rate. It is worth noting that there is yet another type of error that is possible when voting, what is called an overvote. An *overvote* error can occur if a voter selects two candidates in a race where only a single candidate should be chosen; DREs and mechanical lever machines prevent this error. There were no instances of this type of error in this study, so it will not be discussed any further. Error rates for each voting method are presented in Table 4.

*Table 4.* Error types by voting method (standard deviations in parentheses).

| | Extra vote | Undervote | Wrong choice | Total error |
|---|---|---|---|---|
| Sequential DRE | .000 (.000) | .002 (.008) | .011 (.027) | .013 (.031) |
| Direct DRE | .002 (.007) | .131 (.242) | .012 (.048) | .145 (.258) |
| Bubble ballot | .000 (.000) | .002 (.009) | .002 (.008) | .004 (.012) |
| Lever machine | .000 (.000) | .006 (.018) | .011 (.024) | .017 (.028) |
| Punch card | .000 (.000) | .000 (.000) | .002 (.009) | .002 (.008) |

Most interesting was the extremely high undervote error rate for the direct DRE, which was about 13%, as opposed to only .2% for the sequential DRE (Table 4). The large disparity between DRE undervote error rates was responsible for several significant effects seen with a 3 (non-DRE voting method) x 4 (information type) x 2 (navigation type) x 4 (education level) x 3 (subjects' three ballots) x 3 (error type) ANOVA on error rates. (The Greenhouse-Geiser method was used to correct for severe sphericity violations.) The interaction between error type and navigation type was reliable, $F(1.05, 17.82) = 4.76$, $p = .04$, as was the interaction between error type and subjects' three ballots, $F(1.19, 20.28) = 4.98$, $p = .03$. There was also a reliable interaction between navigation type and education, $F(1, 17) = 6.65$, $p = .02$, as well as a four-way interaction between error type, subjects' three ballots, navigation type, and education, $F(1.19, 20.28) = 5.70$, $p = .02$.

The high undervote error rate for the direct DRE also caused several significant main effects. Not surprisingly, the main effect of navigation type on error rates was reliable, $F(1, 17) = 7.39$, $p = .02$), as was the main effect of subjects' three ballots, $F(1.15, 19.61) = 6.64$, $p = .02$. There was also a significant main effect of error type, $F(1.05,17.82) = 14.33$, $p = .001$, which arose because the overall undervote error rate was significantly higher than the extra vote and wrong choice error rates. The abnormally high undervote error rate for the direct DRE can be explained by the number of people who cast their ballots prematurely with the direct DRE, an effect that is discussed more thoroughly in the following section.

It is important to know whether any particular ballot type is significantly more likely to generate ballots containing at least one error. Of greatest interest for this study, one would like to know whether the sequential or direct DRE was more likely to generate such a ballot. An ANOVA comparing the frequency of ballots with at least one error by navigation type found that neither the sequential nor the direct DRE was more likely to generate such a ballot. A similar ANOVA did not show any of the three non-DRE voting methods to be more likely to generate a ballot with at least one error.

### *Casting ballots too soon versus not at all*

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, or a voter can cast their vote prematurely, by pressing the “Record Vote” button too soon. 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. While this raises significant privacy issues, the vote would nonetheless be counted. In this study, if a participant failed to cast their ballot, we cast it for them and counted their choices as intended. However, 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 DRE, four people failed to cast their vote (12.5%), and eight people cast their vote too soon (25%). The difference in “failed to cast” error rates between the two DRE types was not reliable. However, the direct DRE had a significantly greater “cast too soon” error rate than did the sequential, $F(1, 62) = 4.33$, $p = .002$.

Interestingly, not one of the eight people who cast their vote prematurely with the direct DRE were actually amongst those participants excluded as outliers, defined as having greater than 15% errors on all three voting methods they used. This is important, as it suggests that such participants were not necessarily error-prone overall; it was not the case that they made numerous errors with each different voting method they used, but rather that they made a particularly severe type of error by casting their ballot prematurely with the direct DRE only. The number of races that were erroneously omitted in this manner varied, and in several cases, no choices at all had been made at the time of early casting.

More specifically, three of the eight participants who cast a vote prematurely with the direct DRE had undervote error rates of fully 100% their first time with the direct DRE. However, these rates dropped to nearly zero their second time with the direct DRE: error rates were exactly zero for two of those three participants, and .04 for the third. Two of the eight participants had undervote error rates of about one their first time with the direct DRE, yet only improved their error rates to about .5 their second time with the direct DRE. This may be in part due to the two-column layout of the direct DRE’s main page: the 27 races on the ballot were split almost equally between two columns (14 races in the left column and 13 in the right). In cases where votes were cast prematurely with a resulting undervote error rate of about a half, participants had not yet made any selections for races in the right-hand column when they erroneously cast their ballot. One participant did this both times he voted with the direct DRE, and another participant did this once but had an error rate of nearly zero the second time. There was a single participant of the eight who was not significantly affected by casting their vote prematurely, as doing so only resulted in an undervote error rate of .04 one time on the direct DRE.

These unique “cast too soon” and “failed to cast” error types were examined further via two logistic regressions. In the first model, “failed to cast” error rates were regressed on navigation type, information type, non-DRE voting method, education, and age (as a continuous variable). The model was significant overall (Chi-square(8, $N = 64$) = 22.43, $p = .004$), accounting for 64% of the variance in predicting whether participants did or did not fail to cast their ballots. Education was the only single variable that had a statistically significant effect on whether participants failed to cast their vote: $b = 2.18$, $p = .05$. Those participants with more education were less likely to fail to cast their ballots. Four of the 15 people (27%)with a high school education or less, failed to cast their ballot. Two of the 27 people (7%) with some college education failed to cast their ballot, whereas nobody with a bachelors or postgraduate degree made this type of error.

In the second model, “cast too soon” error rates were regressed on the same five variables: navigation type, information type, non-DRE voting method, education, and age. This model was also statistically significant (Chi-square(8, $N = 64$) = 23.12, $p$ = .

003), accounting for 57% of the variance in predicting whether participants cast their ballots prematurely or not. Although the overall model was significant, no individual variable had a significant effect on whether participants cast their ballots too soon.

### *Intentional undervotes*

In a real election, when no choice is indicated for any given race, it is impossible to discern whether such 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 (discussed in previous sections) and intentional undervotes. It was only for the undirected information condition that intentional undervote rates were examined. Such rates would not have been meaningful for any of the directed information conditions, in which participants were not allowed to make their own decisions regarding abstentions.

There was a large disparity in intentional undervote rates between the sequential and direct DREs. While it was quite rare for someone to abstain from a race with the sequential DRE, participants using the direct DRE abstained from nearly half the races on the entire ballot (Figure 18). This main effect of navigation type was statistically significant, $F($1, 2) = 11250.00, $p$ < .001, as seen with a 3 (non-DRE voting method) x 2 (navigation type) x 2 (median split on age) x 4 (education level) x 3 (subjects' three ballots) ANOVA.

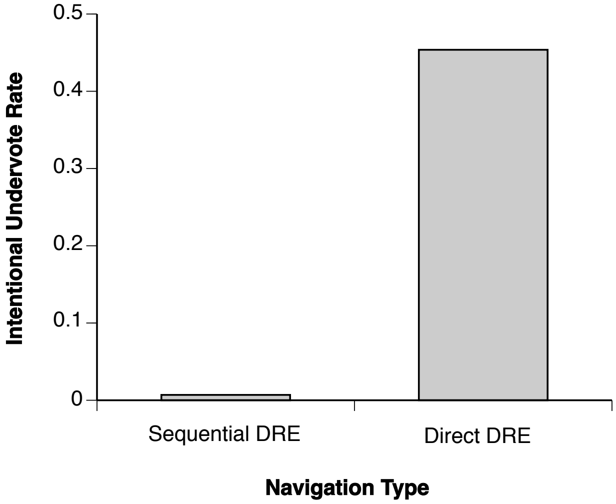


*Figure 18.* Main effect of navigation type on intentional undervote rates in the undirected information condition.

There was also a reliable main effect of non-DRE voting method on intentional undervote rates, *F*(2, 2) = 3256.95, *p* < .001 (Figure 19).

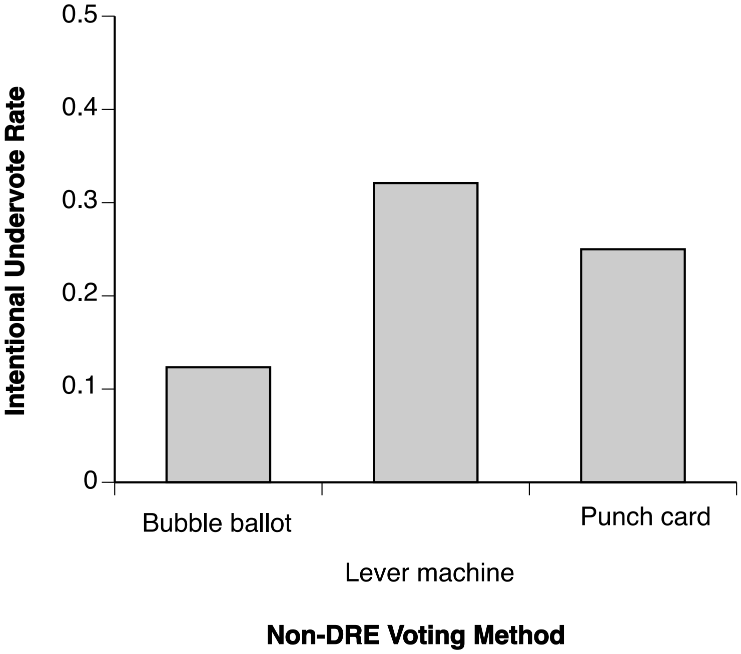


*Figure 19.* Main effect of non-DRE voting method on intentional undervote rates in the undirected information condition.

As is evident in Figure 20, those participants with the least amount of education in the undirected information condition abstained from voting in the greatest number of races. This main effect of education level was statistically reliable, *F*(2, 2) = 3707.73, *p* < .001.

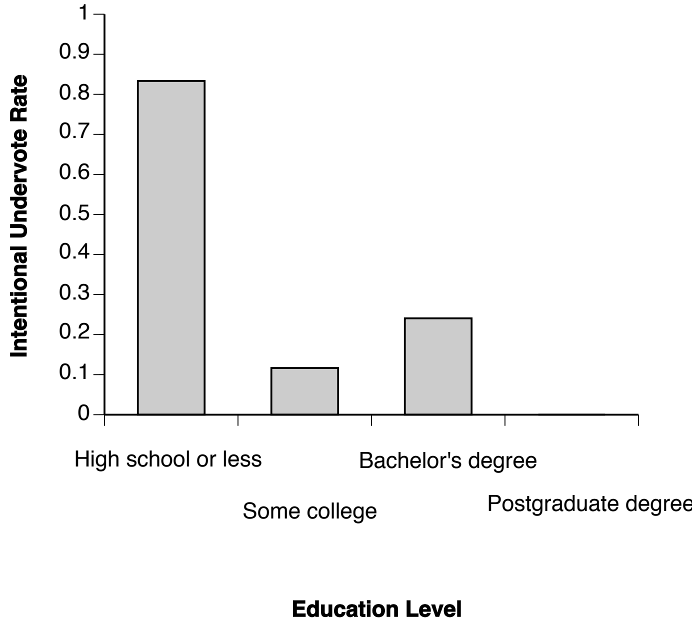


*Figure 20.* Main effect of education level on intentional undervote rates in the undirected information condition.

The median split on age in Figure 21 clearly shows that younger participants (classified as those under the age of 51) abstained from voting in significantly more races than did older participants (those over the age of 51). This main effect of age on intentional undervote rates was statistically reliable, *F*(1, 2) = 16614.20, *p* < .001.

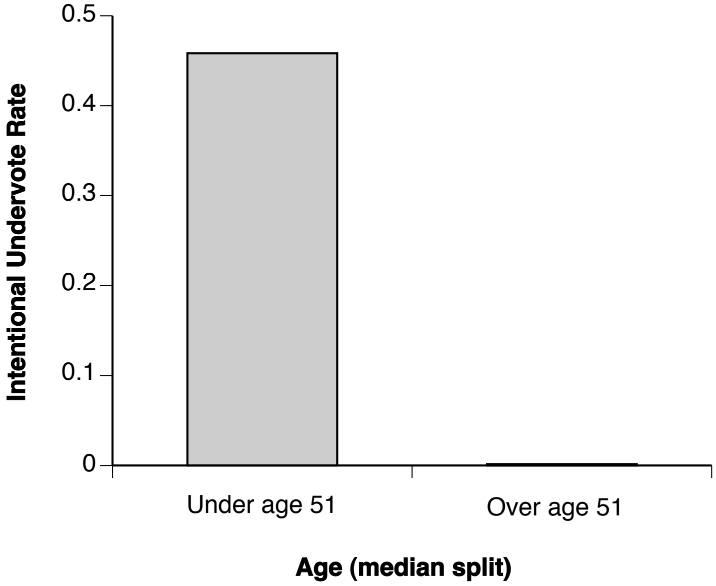


*Figure 21.* Main effect of age (median split) on intentional undervote rates in the undirected information condition.

*Residual vote*

Of 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 rate reflects the total number of votes for any given race that cannot be counted, and is commonly used in Political Science field studies as an indirect measure of accuracy. The residual vote is comprised of overvote errors, undervote errors, and intentional undervotes. The residual vote does not include any information about wrong choice errors because it is impossible to identify such errors without knowing voter intent, which is impractical to do in real elections due to privacy concerns. For the same reason, the residual vote does not differentiate between undervote errors and intentional undervotes; although an intentional abstention is clearly not an error, in both cases no vote can be counted for the race in question. However, because the design of this study explicitly tracked voter intent, it was possible to accurately identify wrong choice errors, as well as to differentiate between undervote errors and intentional undervotes.

To compute what would have been reported as the residual vote for this study, overvotes, undervotes, and intentional undervotes were summed. This 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 overvotes, undervotes, and wrong choice errors combined (Figure 22). In our studies, the undirected information condition is clearly most ecologically valid, as voters in real-world elections have the freedom to abstain from voting or select their own candidates based on personal preference. If the correlation between the residual vote and “total” error rate (for the undirected information condition) is not significant, it would suggest that the real-world residual vote is an inflated measure of error. While this is intuitive, comparisons between direct versus indirect measures of error have not been reported (ours are the first laboratory studies we know of that measured true error rates). These data and results are therefore important, as they provide empirical support that what has previously been only assumed is likely true.

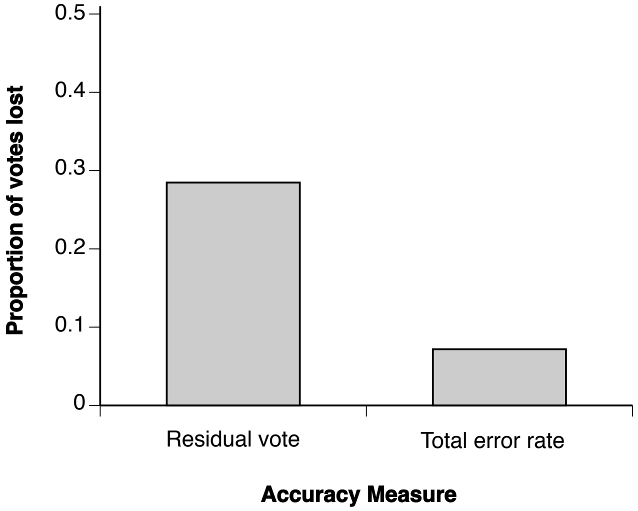


*Figure 22.* Residual vote rates versus true error rates for the undirected information condition only.

The overall residual vote rate did not correlate reliably with the overall true error rate, nor were there significant correlations between the residual vote rate and true error rate for the DRE or non-DRE methods when considered separately. A 3 (non-DRE voting method) x 4 (information type) x 2 (navigation type) x 4 (education level) x 3 (subjects' three ballots) ANOVA examined effects of these five variables on residual vote rates. The only statistically reliable result was the main effect of subjects' three ballots: $F(1.05, 5.24) = 9.44$, $p = .03$. (The Greenhouse-Geiser method was used to correct for a severe sphericity violation.) As one would expect based on results from the prior efficiency and effectiveness analyses, the main effect of subjects' three ballots was again explained by a reliable up-down-up pattern in the data: $F(1, 5) = 8.53$, $p = .03$.

*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. Overall, SUS ratings for the alternate methods as a whole were much lower than ratings for the DREs, and DRE ratings did not differ depending on whether state information was present. The sequential DRE received the highest mean SUS score of any voting method, and participants were much less variable in rating their satisfaction with that interface than with any other voting method (Figures 23 and 24).

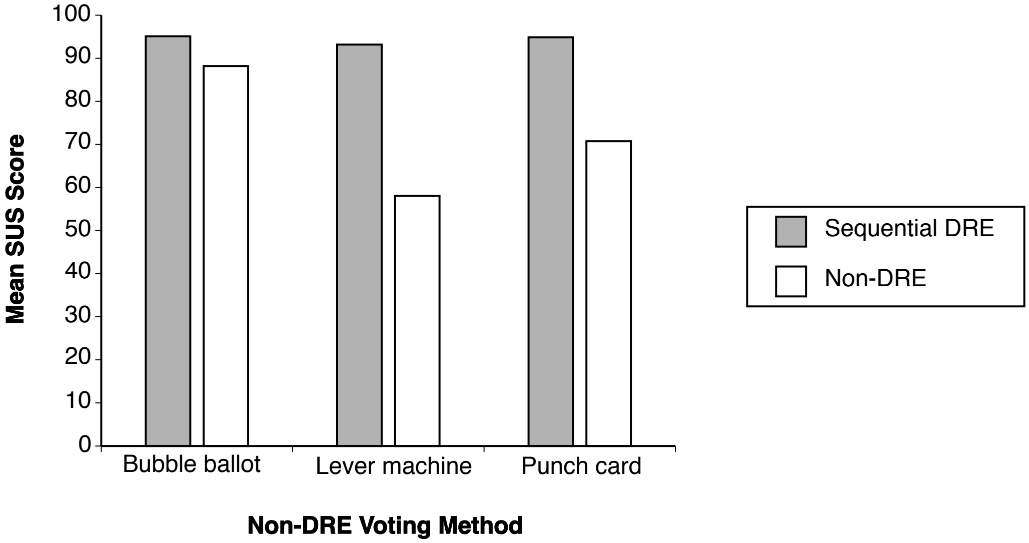


*Figure 23.* Mean SUS scores for sequential DRE versus non-DRE voting methods.

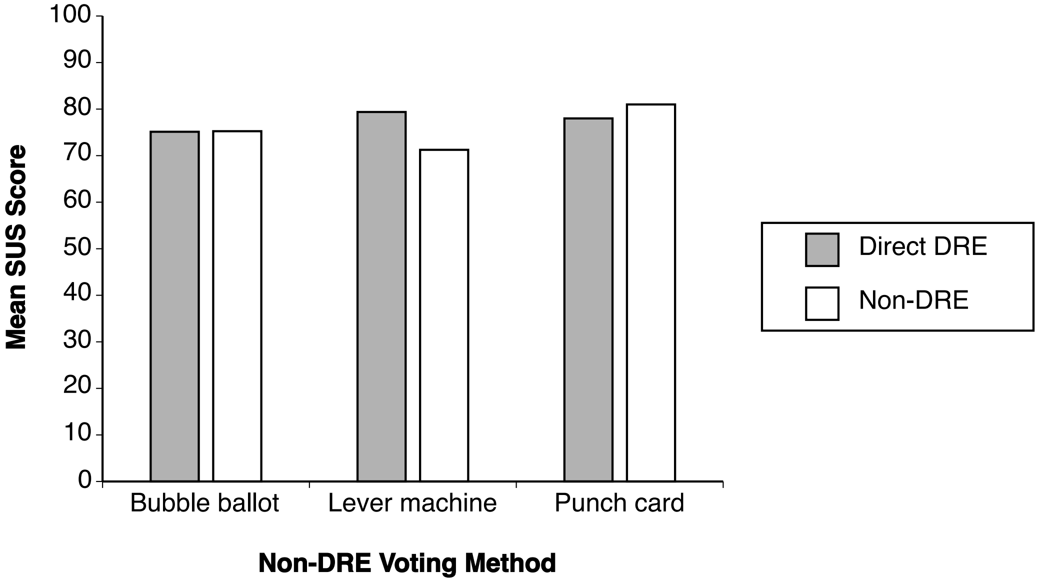


*Figure 24.* Mean SUS scores for direct DRE versus non-DRE voting methods.

A 3 (non-DRE voting method) x 4 (information type) x 2 (navigation type) x 3 (subjects' three ballots) ANOVA on SUS scores revealed several statistically significant effects. Since the sequential DRE consistently received much higher SUS scores than did the direct DRE, it was not surprising that the main effect of navigation type on SUS scores was reliable: *F*(1, 38) = 9.53, *p* = .004. There was also a main effect of subjects' three ballots on SUS scores, *F*(2, 76) = 4.64, *p*<.001. This was due to the extremely high sequential DRE SUS scores and the fact that participants voted first with a DRE, then with an alternate method, and then again with a DRE, which resulted in a reliable high-low-high effect in the SUS data, *t*(61) = 3.34, *p* = .001. This high-low-high pattern of SUS scores also played a role in the reliable interaction between subjects' three ballots and navigation type, *F*(2, 76) = 8.79, *p*<.001. Only when the navigation type was sequential was this pattern statistically reliable (decomposed with a posthoc interaction contrast, *F*(1, 60) = 8.95, *p* = .004). In other words, while SUS scores for the sequential DRE were significantly greater than for the corresponding non-DRE voting methods, SUS scores for the direct DRE could not be similarly distinguished from the other methods.

There was also a reliable interaction between subjects' three ballots and non-DRE voting method, *F*(4, 76) = 2.64, *p* = .04. Simple main effects tests showed that only when the non-DRE voting method was the lever machine was there a reliable effect of subjects' three ballots on SUS scores. In other words, the significant high-low-high effect of subjects' three ballots on SUS scores was not present when the alternate method being used was the bubble ballot or punch card.

*Discussion, Study 2*

As in Study 1, there were no reliable effects of state information on efficiency, effectiveness, or satisfaction, so it will not be discussed any further. The most important manipulation in Study 2 was the addition of a new DRE, one with a direct access navigation model. This single interface modification of changing the DRE navigation style from sequential to direct resulted in a myriad of significant effects on both objective and subjective usability; intentional undervote rates, efficiency, effectiveness, and satisfaction were all greatly affected by the new interface design.

When participants were given the freedom to make their own choices in the undirected information condition, the difference in rates of intentional abstention between the two DRE types was dramatic. When required to page through every race with the sequential DRE, voters almost never abstained from voting. Conversely, when offered the opportunity to avoid races and skip straight to casting a ballot with the direct DRE, voters chose to abstain from voting in nearly half of all races. As one would expect, choosing to vote in far fewer races made the direct DRE significantly faster than the sequential in the undirected information condition.  Overall, participants voted more quickly with the non-DRE methods than with the DREs, likely due to the extremely long ballot completion times for the sequential DRE in the undirected information condition.

Not only did the direct DRE have a much higher *intentional* undervote rate than did the sequential, but its undervote *error* rate was significantly greater as well. With the direct DRE, one fourth of participants cast their ballot prematurely, whereas no participant made this type of mistake with the sequential DRE. Casting a ballot prematurely is a particularly serious error, as it irrevocably disenfranchises a voter. Another major error occurs when voters fail to cast their ballot at all. This type of postcompletion error is an issue in real-world elections and has been termed the “fleeing voter” problem.

It seems likely that participants were sensitive to the poor effectiveness of the direct DRE, as its SUS scores were significantly lower than those for the sequential DRE. As in Study 1, the sequential DRE was by far the superior interface in terms of subjective usability. However, it was actually no better than any of the other voting methods in terms of efficiency or effectiveness. The effectiveness of voting systems has been studied indirectly in real-world elections via the residual vote. Interestingly, when this study’s residual vote rate was compared with its true error rate, there was little agreement between the two measures. A direct calculation of error rates would clearly be preferable over an indirect method, but without compromising voters' privacy, there does not seem to be a viable way to replace the residual vote with a more direct measure of accuracy. This by no means implies that the residual vote does not provide useful information, but merely suggests that it might not tell the whole story.

### *General conclusions and future directions*

When HAVA was passed in 2002, it was widely assumed that DREs were superior to the older technologies they were replacing despite the absence of any research supporting that assumption. Furthermore, there remains a dearth of systematic research within the large design space offered by DREs. The two studies reported here begun to address this lack of research, and found that the assumption that DREs are superior is unfounded, despite voter preferences. Across both studies, participants were neither faster nor less error-prone with the DREs relative to the other methods. Nonetheless, they consistently found the sequential DRE significantly more satisfying, an interesting disassociation between preference and performance. In fact, with an average SUS score above 90, the sequential DRE would be considered a "truly superior product" according to recent research evaluating a broad range of technologies (Bangor, Kortum, & Miller, in press).

Study 2 is the first in which a DRE design feature (navigation style) has been directly manipulated and significantly impacted *both* subjective and objective usability. Changing the navigation style from sequential to direct had deleterious impacts on satisfaction and effectiveness. Only in the undirected information condition was the direct DRE any faster than the sequential, yet this was accompanied by reduced voter participation, since intentional abstentions were much more frequent with the direct DRE. Most importantly, with the new direct access navigation system, numerous voters were disenfranchised after casting their ballots prematurely.

There are relatively simple interface changes that should be examined in future research. Take for instance the newly discovered error of premature ballot casting in Study 2. The screen location of the "Record Vote" button was likely a contributing factor, as it was located in the exact same place as the "Next Page" button was, at the bottom right-hand corner of the screen (Figure 6). The two buttons were also visually similar in terms of size, shape, and color. Future research could examine effects of changing the "Record Vote" button's location and manipulating its visual characteristics to make it easily distinguishable from the "Next Page" button. It may be that such simple interface changes could significantly reduce the frequency of errors such as premature ballot-casting, as well as help with the "fleeing voter" problem. Additionally, future research could examine a hybrid of the sequential and direct DRE, where voters still page sequentially through all races on the ballot, but have the option to skip straight to casting their ballot at any time. Outcomes of elections can, and have been, close enough that even small reductions in error rates could have huge societal impact.

While DREs are often touted as being more accessible for voters with disabilities, there is little if any research supporting such claims. In fact, recent findings indicate that commercial DREs are not especially successful at addressing accessibility issues (Runyan, 2007). Future research should investigate ways in which the flexibility of computer-based systems can be better harnessed to improve DRE usability for those voters with disabilities. For example, support for multiple font sizes may benefit voters with low vision, while auditory interfaces have the potential to aid fully non-sighted individuals.

While some results from these studies pertain mainly to DREs and U.S. elections, this research nonetheless offers important design lessons that apply to system interfaces of any type and in any culture. In both studies, a disassociation between preference and performance was found; the possibility for a disconnect between subjective and objective usability is something we all need to be aware of as designers and human factors professionals. Although it is not always the case, some design problems may require a tradeoff between these two facets of usability. The relative importance of subjective versus objective usability should be weighed carefully before expending valuable resources implementing new technologies on a large scale. As seen in these studies, there is no guarantee that new systems are necessarily better than those they are intended to replace.

Usability and human factors have only recently begun to receive widespread attention in the voting arena; hence data are still scarce, both from the laboratory and from the field. Therefore the two studies reported here contribute much-needed data from a well-controlled laboratory setting. These data serve as an excellent complement to large-scale field studies that have been previously conducted. Usability in voting systems is a topic that is steadily gaining both popular and scientific interest; as such, these findings may even have the potential to reach the ears of a broader audience than solely academic researchers.

## REFERENCES

Alvarez, R. M., & Alsolabehere, S. (2002). California votes: The promise of election day registration. Demos: A network for ideas and action. Available from www.demos-usa.org

Alvarez, R. M., Goodrich, M., Hall, T. E., Kiewiet, D. R., & Sled, S. M. (2004). The complexity of the California recall election. Available from www.apsanet.org

Ansolabehere, S. (2001). Prepared Remarks for the House Science Committee. Retrieved August, 2005, from http://www.house.gov/science/full/may22/ansol.htm

Ansolabehere, S., & Stewart, C., III. (2005). Residual votes attributable to technology. The journal of politics, 67, 2, 365-389.

Bangor, A., Kortum, P. T., & Miller, J. T. (in press). An empirical evaluation of the System Usability Scale (SUS). To appear in International Journal of Human-Computer Interaction.

Benway, J. P., & Lane, D. M. (1998). Banner blindness: Web searchers often miss "obvious" links. Internetworking, 1.3. Online newsletter of the Internet Technical Group: http://www.internettg.org/newsletter/dec98/banner_blindness.html.

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.

Bullock, C. S. III, & Dunn, R. E. (1996). Election roll-off: A test of three explanations. Urban Affairs Review, 32, 71-86.

Byrne, M. D., Greene, K. K., & Everett, S. P. (2006). Usability of voting systems: Baseline data for paper, punch cards, and lever machines. Proceedings of CHI 2006. Montreal, Quebec, Canada.

Caltech/MIT Voting Technology Project. (2001). Voter registration. Available from http://www.vote.caltech.edu/reports

Caltech/MIT Voting Technology Project. (2004a). Immediate steps to avoid lost votes in the 2004 presidential election: Recommendations for the Election Assistance Commission. Available from http://www.vote.caltech.edu/reports

Caltech/MIT Voting Technology Project. (2004b). Insuring the integrity of the electoral process: Recommendations for consistent and complete reporting of election data. Available from http://www.vote.caltech.edu/reports

Caltech/MIT Voting Technology Project. (2004c). On the discrepancy between party registration and presidential vote in Florida. Available from http://www.vote.caltech.edu/reports

Caltech/MIT Voting Technology Project. (2004d). Seven steps to make sure your vote is counted: A guide for American Voters. Available from http://www.vote.caltech.edu/reports

Caltech/MIT Voting Technology Project. (2004e). Voting machines and the underestimate of the Bush vote. Available from http://www.vote.caltech.edu/reports

Caltech/MIT Voting Technology Project. (2005a). Public attitudes about election governance. Available from http://www.vote.caltech.edu/reports

Caltech/MIT Voting Technology Project. (2005b). Testimony on voter verification: Presentation to Senate Committee on Rules and Administration. Ted Selker, MIT. Available from http://www.vote.caltech.edu/reports

Caltech/MIT Voting Technology Project. (2005c). Voter registration: Past, present, and future. Michael Alvarez, Caltech. Available from http://www.vote.caltech.edu/reports

Center for American Politics and Citizenship, & Human-Computer Interaction Lab. (2006). A study of vote verification technology conducted for the Maryland State Board of Elections Part II: Usability study. University of Maryland, College Park, MD, 20742. Available from www.capc.umd.edu/rpts

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. Santa Monica, CA: Human Factors and Ergonomics Society.

Frokjaer, E., Hertzum, M., & Hornbaek, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? CHI, 345-352.

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. Vancouver, BC, Canada.

Herrnson, P. S., Bederson, B. B., & Abbe, O. G. (2002). An evaluation of Maryland's new voting machine. The Center for American Politics and Citizenship and Human Computer Interaction Lab, University of Maryland, College Park, MD 20742. Available from http://www.capc.umd.edu/rpts

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

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

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

Kimball, D. C., Owens, C. T., & Keeney, K. M. (2003). Unrecorded votes and election reform. Spectrum: The journal of state government. 34-37.

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.

Mebane, W. (2004). "The Wrong Man is President! Overvotes in the 2000 Presidential Election in Florida." Perspectives on Politics, 2, 4, 525-533.

Runyan, N. (2007). Improving access to voting: A report on the technology for accessible voting systems. Retrieved April 21, 2008 from http://demos.org/pubs/improving_access.pdf.

Sinclair, D. E., & Alvarez, R. M. (2004). Who overvotes, who undervotes, using punchcards? Evidence from Los Angeles County. Political Research Quarterly, 57, 15-25.

Tomz, M., & Van Houweling, R. P. (2003). How does voting equipment affect the racial gap in voided ballots? American Journal of Political Science, 47, 1, 46-60.

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

United States Government, Election Assistance Commission. (2005). Voluntary Voting System Guidelines. 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, 793-84.

Wattenberg, M. P., McAllister, I., & Salvanto, A. (2000). How voting is like taking an SAT test: An analysis of American voter rolloff. American Politics Quarterly, 28, 234-50.