ABSTRACT

Effects of Multiple Races and Header Highlighting on Undervotes in the 2006 Sarasota General Election: A Usability Study and Cognitive Modeling Assessment

by

Kristen K. Greene

Large-scale voting usability problems have changed the outcomes of several recent elections. The 2006 election in Sarasota County, Florida was one such incident, where the number of votes lost was nearly 50 times greater than the margin of victory for the US Representative race. Multiple hypotheses were proposed to explain this incident, with prevailing theories focused on malicious software, touchscreen miscalibration, or poor ballot design. Study 1 aimed to empirically determine whether Sarasota voters unintentionally skipped the critical US Representative race due to poor ballot design. The Sarasota ballot was replicated initially, then header highlighting and number of races presented on the first screen were manipulated. While the presentation of multiple races had a significant effect on undervotes in the US Representative race, header highlighting did not. Nearly 20% of all voters (27 of 137) skipped the race their first time on that screen, an even greater undervote rate than that originally seen in Sarasota. In conjunction with other research, Study 1 results strongly suggests that the 2006 Sarasota election was almost certainly a human factors problem. A cognitive model of human voters was developed based on Study 1 data. Model predictions were then compared with behavioral data from Study 2, in which participants voted on a replica of the Charlotte County, Florida 2006 ballot.

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, Dr. David M. Lane, and Dr. Dan S. Walach, for their invaluable guidance and feedback. Finally, many thanks to my family and friends for their inexhaustible encouragement. Last but not least, a special thanks to Sarah Everett, Bryan Campbell, and Frank Tamborello.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

After the Help America Vote Act was passed in 2002, traditional voting systems (paper ballots, punch cards, and lever machines) were largely replaced with new direct-recording electronic voting machines (DREs) (United States Government [US Gov.], 2002). It was widely assumed that DREs would be more usable than the technologies they were replacing, yet usability problems with DREs in recent elections have been of even greater magnitude than usability issues seen previously with older voting methods. The 2006 election in Sarasota County, Florida was one such incident, where the number of votes lost was nearly 50 times greater than the margin of victory for the US Representative race. Although large-scale voting usability problems have changed the outcomes of several recent elections, existing accounts of usability failures in voting only address *what* happened when voters were confronted with various poor ballot designs, while explanations of *why* certain design features consistently result in specific types of voter errors are lacking. Furthermore, voting usability issues are currently addressed only *post*-election. In the future, a means of assessing usability *pre*-election is needed, to help prevent the reoccurrence of large-scale usability problems similar to those seen in recent elections. In order to achieve this long-term goal, a more thorough knowledge of how various ballot design characteristics affect voter performance is necessary. The current studies evaluated several ballot design principles, in an effort to help differentiate between multiple hypotheses proposed to explain a recent DRE usability problem from the 2006 Sarasota election. In Study 1, human participants voted on a ballot design analogous to the problematic real-world original. Based on Study 1 data, a cognitive

model was constructed by synthesizing knowledge of general usability design principles with knowledge of human attention and perception. The model was then used to predict effects on human performance expected with novel interface changes. Human participants voted on ballots with those identical interface changes in Study 2, and model predictions were compared to human data.

In general, methodology for the current studies was heavily based on a previous series of experiments. Therefore, the following literature review begins with a summary of relevant experimental studies that directly measured voting system usability. Subsequent sections review field studies that examined usability via indirect methods (i.e. the residual vote), with special emphasis on the ballot design and competing hypotheses from the election that motivated these studies.

Direct evaluation of voting system usability: Experimental data

In order to determine whether new electronic voting systems actually offer usability improvements over traditional voting methods, it is necessary to have baseline usability data for both the newer and older technologies. In the first of a series of studies, Everett, Byrne, and Greene (2006) began to address this need for baseline data, by comparing the usability of three different styles of paper ballots: open response ballots, bubble ballots, and arrow ballots. Usability was assessed via ballot completion times, error rates, and participants' responses to a standardized satisfaction questionnaire, the System Usability Scale (SUS) (Brooke, 2006). These measures addressed the recommendation by the National Institute for Standards and Technology (NIST) that

usability be evaluated using three metrics suggested by the International Organization for Standardization (ISO): efficiency, effectiveness, and satisfaction (Laskowski et al., 2004). Neither efficiency (ballot completion time) nor effectiveness (error rate) were affected by the type of paper ballot used in the 2006 study by Everett et al. However, voting method did have a significant effect on satisfaction, with the bubble ballot receiving the highest SUS scores of the three paper ballots. The open response and arrow ballots could not be distinguished from one another in terms of subjective preference. While error rates were not significantly different depending on which type of paper ballot was used, over 11% of ballots completed contained at least one error. Such high error rates were disturbing, given that participants in this study were comprised of Rice University undergraduates, an extremely intelligent and highly educated sample.

In a subsequent study, Greene, Byrne, and Everett (2006) used participants more representative of the voting public, by sampling from the general population in Houston, Texas. This study examined the usability of mechanical lever machines, as well as two of the three previously studied paper ballots: the bubble and arrow style ballots. Although voting method again had no significant impact on efficiency or effectiveness, nearly 16% of all ballots cast contained at least one error. Again, voting method had a marked impact on satisfaction. Participants were by far most satisfied with the bubble ballot, and least satisfied with the lever machine. In a later study, Byrne, Greene, and Everett (2007) added the use of punch cards to their voting usability research. Consistent with prior results, the bubble ballot again received significantly higher SUS scores than did the lever machine, punch card, or arrow ballot. The percentage of ballots containing at least one

error was higher than in previous studies (26%), although no voting method was more or less prone to generating errors.

In 2008, Everett and colleagues conducted their first study directly comparing the usability of older voting systems with newer electronic voting technology, by evaluating bubble ballots, lever machines, punch cards, and a new prototype DRE. Differences in efficiency and effectiveness were small between the various voting methods. However, the DRE elicited significantly higher SUS scores than did any of the older voting methods. As in previous studies, the bubble ballot again received high SUS scores, but it was no longer participants' favorite once they voted with the new DRE. While the DRE enjoyed a large advantage in terms of subjective usability ratings, it did not have any corresponding objective performance benefits. Such a disassociation between subjective and objective usability is not uncommon (Frokjaer, Hertzum, & Hornbaek, 2000).

While Everett and colleagues (2008) used a prototype DRE developed in-house, other researchers have used commercially available DREs. Herrnson and colleagues (2008) included a variety of commercial DREs in a large-scale usability evaluation of six electronic voting systems and four verification systems. They used several different methods of evaluation: expert reviews, a laboratory experiment, and a field study. In their laboratory experiment, 42 participants voted on six different DREs: a Diebold Accuvote-TS, an ES&S Model 100, a Hart InterCivic eSlate, an Avante Vote-trakker, a Nedap Liberty Vote, and a Zoomable prototype developed specifically for the authors. The ballot used fictional candidate names, and contained 18 offices and four ballot questions. Participants were given a voter guide and asked to circle their choices and keep it with

them while they voted, but were instructed who to vote for in some races, and were asked to do a write-in, change a vote, and intentionally omit a down-ballot race. Allowing people to make their own selections, but then telling them to change those selections in certain ways, was a slightly odd aspect of the study methodology. Another methodological issue was the study's use of a non-standardized questionnaire to evaluate subjective usability. Had voter satisfaction been measured via a standardized instrument, such as the SUS, it would have facilitated more accurate comparison across future studies and different methodologies. One final concern with the Herrnson et al. (2008) study is the lack of efficiency data for any of the voting systems; if ballot completion times were recorded, they were not published.

Despite these limitations, the Herrnson et al. (2008) study was a milestone in voting system usability research. When voters were not asked to perform any special tasks and were given a standard office-bloc ballot design, ballots were cast accurately over 97% of the time, with few differences between voting systems. However, when voters were asked to do any special tasks (change a choice, vote straight-party ticket, or select multiple candidates), accuracy dropped noticeably, down to 80-90%. Furthermore, sharp differences emerged between voting systems, with the Avante's accuracy rate at only 50%, and the Hart InterCivic's rate at 72%. Approximately 20% of voters cast ballots that were not entirely accurate. Voters seemed insensitive to these objective usability differences; they rated all six DREs favorably and had high confidence with the touch screen systems, even judging them as more trustworthy than paper. This disassociation between objective and subjective usability with DREs is similar to

previous findings by Everett and colleagues using a prototype DRE (2008).

The ability to measure both objective and subjective usability has been a major contribution of laboratory experiments in the voting arena. Despite NIST's recommendation that usability be assessed via metrics of efficiency, effectiveness, and satisfaction, data from studies that directly measured all three aspects of usability remain relatively scarce. Indirect measures of usability are much more common, but they are generally limited to a single facet of usability, that of effectiveness. Although use of the residual vote as an indirect measure of accuracy (i.e. effectiveness) is a long-standing tradition in the Political Science domain, the residual vote has not traditionally been discussed explicitly as a method for assessing "usability" per se. Nonetheless, this indirect measure has consistently been a viable means of drawing attention to usability problems post-election, as described in the following section.

Indirect evaluation of voting system usability: Field data on the residual vote

Given the decentralized nature of election administration in America, election policies and procedures can vary widely between jurisdictions, with voting systems and ballot designs that differ both between and within states. When a given state has neighboring counties comprised of voting populations with similar characteristics (socioeconomic status, political preference, ethnicity, etc.), one would expect election results to be similar between counties. When large and unexpected differences in election results occur between neighboring counties, or between a single county in comparison to its state, it causes voters and candidates alike to question the legitimacy of the election.

This has happened repeatedly in recent years, which has greatly increased national interest in voting systems. Much of this public interest has been focused on auditability and potential security vulnerabilities of newer direct recording electronic voting machines (DREs), with little focus on usability and human factors. This is surprising, given that poor ballot design, rather than failed or compromised technology, has been the most likely cause of aberrant and questionable results in numerous elections. In several notable cases, the number of votes lost due to poor ballot design far outnumbered the margin of victory.

The number of lost or unrecorded votes, termed the residual vote rate, is an indirect measure of voter error commonly used in Political Science. It is typically calculated as the difference between voter turnout and the number of valid votes actually cast for any given race, and is used to compare accuracy between different voting methods (Kimball & Kropf, 2005). The residual vote is comprised of both undervotes and overvotes. Undervoting occurs if a voter makes no selection at all for a given race. An undervote can be either intentional or unintentional. Many voters plan only to vote in the presidential race and then skip the remaining down-ballot races; their down-ballot undervotes would be intentional omissions rather than errors. On the other hand, numerous voters plan to make a selection in every race, making their undervotes true errors. Without a way to ascertain voter intent, one cannot disentangle intentional omissions from undervote errors. In contrast to an undervote, an overvote occurs if a voter selects more than the allowed number of options for a particular race. For example,

choosing more than one candidate for president would mean neither vote was recorded for that race.

Residual vote rates of greater than one percent are rare for “top-of-the-ticket” races, especially for president. Although the residual vote is not a direct measure of voter error, unusually high residual vote rates have repeatedly proven useful in elucidating problems post-election. The Brennan Center for Justice (2008) examined ballots from counties with abnormally high residual vote rates, and found those counties had inevitably used poorly designed ballots in comparison to their neighboring county or their state. The Brennan Center report presents several general ballot design problems, supported by specific examples of ballots used and their corresponding election outcomes. Likely the most well-known of those incidents detailed in the report was the use of the “butterfly ballot” in the 2000 general election in Palm Beach County, Florida. Effects of the “butterfly ballot” have been discussed prior to the Brennan Center report (Wand et al., 2001), and this incident may arguably be one of the most publicly memorable large-scale examples of a design error that resulted in reduced usability for thousands of voters.

The butterfly ballot used a single column of punch holes flanked by two columns of candidate names for the same office (Figure 1).

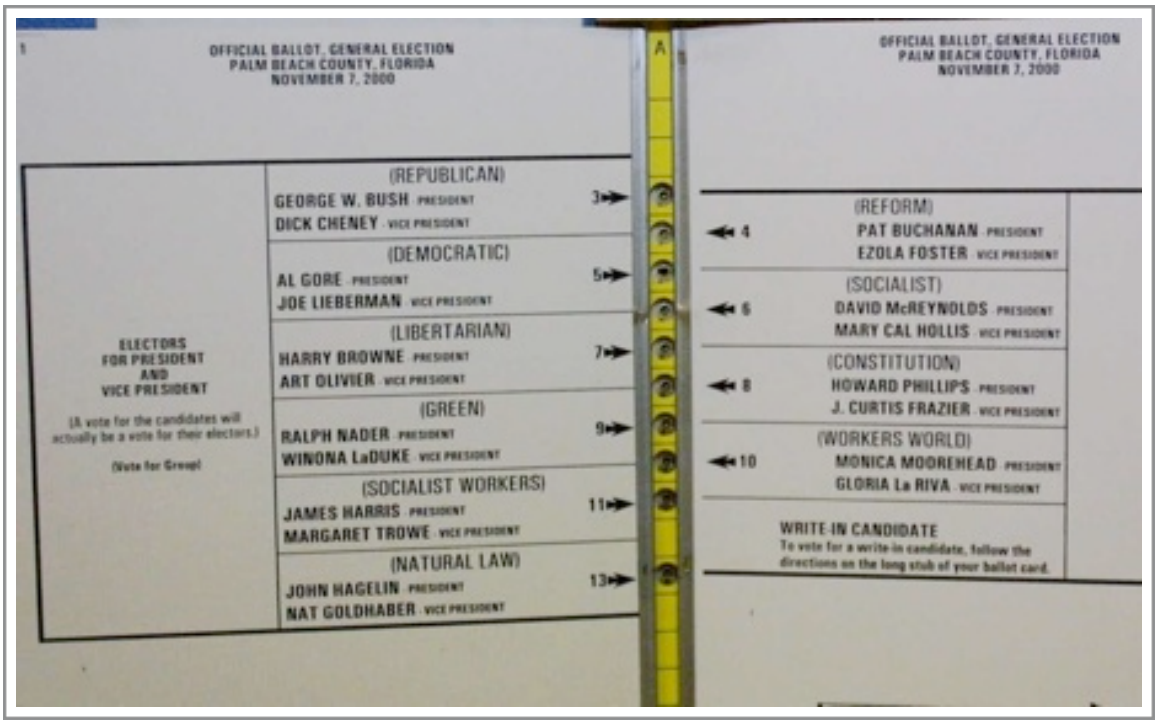


Figure 1. The "butterfly ballot" used in Palm Beach County, Florida, 2000.

The 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 have perceived that punching the first hole corresponded to a vote for Bush, and that punching the second hole corresponded to a vote 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.

Indeed, it was estimated that over 2,000 Democratic voters mistakenly selected Buchanan when using the "butterfly ballot." Buchanan received 3,411 votes in Palm Beach County, which was over three times the number he received in other Florida counties. The residual vote rate in Palm Beach County was 6.2%, whereas the residual vote rate for the state was only 2.9%. There were 28,746 residual votes in the county, and the margin of victory in the state was only 537 votes (Brennan Center for Justice, 2008). Significant statistical evidence suggests that if precinct-tabulated optical scan ballots had been used throughout Florida, Al Gore would have been elected president rather than George Bush, with a margin of victory of over 30,000 votes (Mebane, 2004).

Palm Beach County was not the only Florida county to split candidates for the same office onto different pages in the 2000 general election. Duval County also used a two-page ballot design that year, splitting the ten presidential candidates across pages by listing five on the first page and five on the second page. Almost 22,000 votes were not counted, because people chose one candidate on the first page, then chose yet another candidate on the second page. This ballot design was used in an area of the state where Democrats had launched an aggressive campaign to register new voters, with voter education materials reminding people to "vote that second place." The residual vote rate in Duval County was 9.3% for president, compared to only 2.9% statewide (Brennan Center for Justice, 2008).

Although the election issues in Palm Beach and Duval Florida counties arose due to splitting candidate names for a single race across multiple pages of a ballot, merely ensuring that all candidate names for a given race are displayed on one page does not

guarantee a well-designed layout. For example, even when a ballot uses only a single page, it can be problematic to split candidates for one race across two columns. This occurred in the 2002 general election in Kewaunee County, Wisconsin, where the paper ballot split the race for governor into two columns. The residual vote rate for that contest was a surprising 11.8% in the county, but only 1.1% statewide. The number of votes lost for that race in Kewaunee County (over 700) was larger than the margin of victory there (only 400 votes), although the state overall was won by a comfortable margin (Brennan Center for Justice, 2008). The finding that greater numbers of unrecorded votes occur with ballots that list candidates for the same office in multiple columns or on multiple pages has been reported previously (Sinclair & Alvarez, 2004).

Optical scan ballots commonly present the voter with multiple columns of races and candidate names on a single sheet, but what is less common and quite problematic is the placement of response options on both sides of candidate names. This ballot design flaw occurred in Hamilton County, Illinois in the 2002 general election (Figure 2).

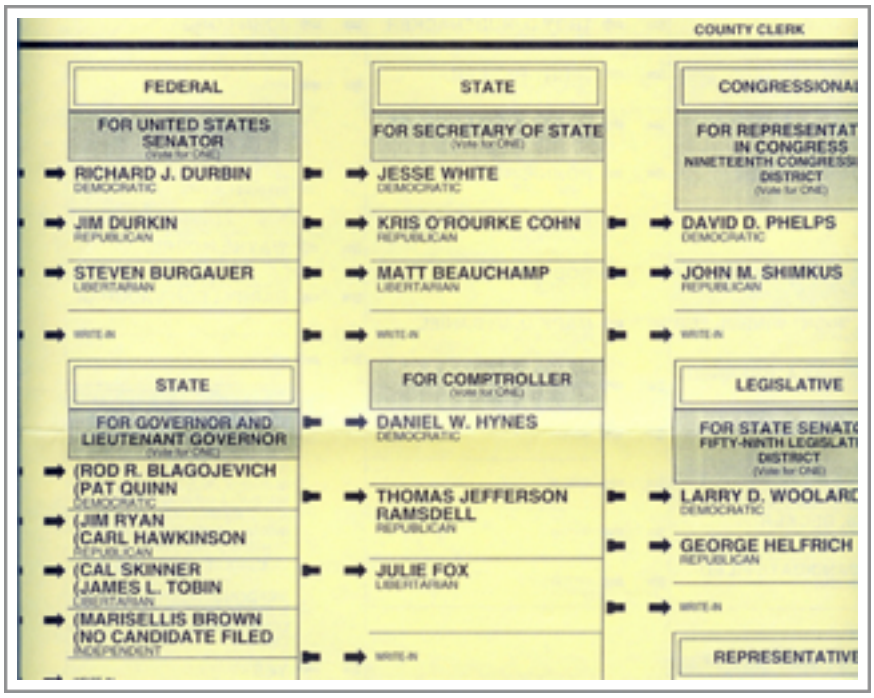


Figure 2. Section of the ballot used in Hamilton County, Illinois, 2002.

The first and second columns of the Hamilton County ballot had races that lined up exactly next to one another. Numerous voters erroneously marked the arrow to the right of a candidate's name rather than to arrow to the left, likely due to the fact that people read from left to right, combined with the limited spacing between columns that made it difficult to tell which arrow corresponded to exactly which candidate name. Two races were affected by the poor ballot design in Hamilton County: the race for U.S. Senate, as well as the race for Governor. The residual vote rate in Hamilton County was 9.3% for the U.S. Senate race, compared to only 4.5% statewide. Similarly, the county's residual vote rate was 5.0% for Governor, compared to 3.1% statewide (Brennan Center for Justice, 2008).

In general, “complete-the-arrow” style ballots have higher residual vote rates than do “fill-the-oval” style ballots (Kimball & Kropf, 2005). This may be in part due to voters’ greater familiarity with the task of darkening a bubble, such as on many standardized tests, than with completing the missing section of an arrow. To compare residual vote rates between complete-the-arrow and fill-the-oval style ballots, the Brennan Center for Justice (2008) studied 641 counties using precinct count optical systems and 767 counties using central count optical scan systems. The Brennan Center found that in the 2004 general election, complete-the-arrow style ballots had a residual vote rate of 0.9% with precinct count optical scan systems, whereas fill-the-oval style ballots had a residual vote rate of 0.6%. The same pattern of poorer performance with complete-the-arrow ballots existed when considering central count optical scan systems as well; with central count systems, the residual vote rate for complete-the-arrow ballots was 2.3%, versus 1.6% for fill-the-oval ballots. It was suggested that the use of central count optical scan machines increased the difference in residual vote rates seen between the two ballot styles because voters do not have an opportunity to correct their errors with central count systems. When central count systems are used, they often tabulate votes cast in different polling places, or those received via absentee ballots.

The Brennan Center for Justice (2008) used a different data set from the 2004 election to illustrate yet another ballot design flaw, that of leaving columns or rows for disqualified candidates. This occurred in Montgomery County, Ohio, when Ralph Nader was disqualified from the presidential ballot. Rather than entirely removing the affected row from the ballot, several counties instead used the words “Candidate Removed” in

place of Ralph Nader's name. This was done because many voting machines had already been programmed to scan ballots that included Ralph Nader as a candidate. Because election laws in Ohio mandate that names of candidates are rotated by precinct, every fifth precinct in Montgomery County, Ohio, used ballots with the "Candidate Removed" format pictured in Figure 3.

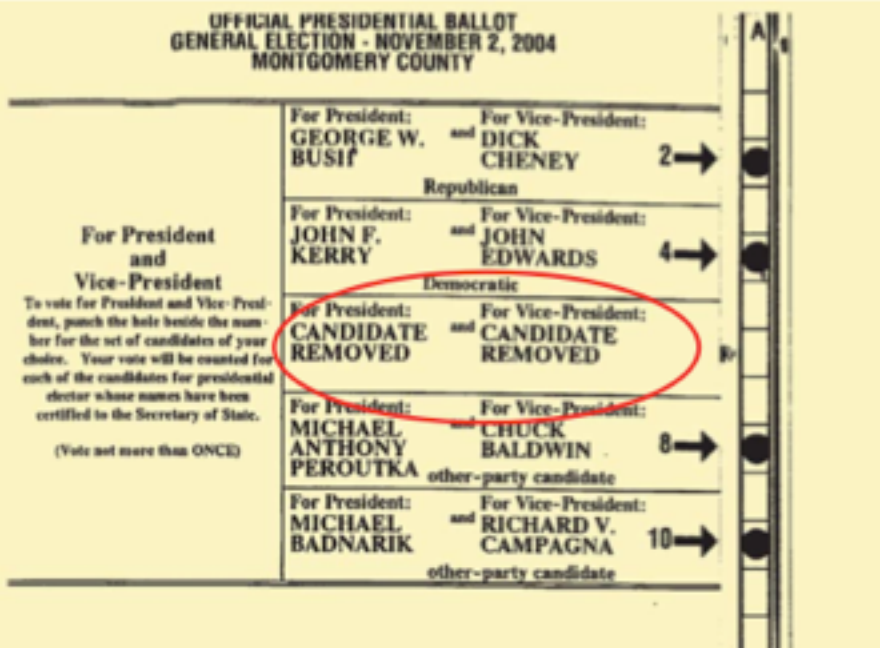


Figure 3. Section from one of several ballot formats used in Montgomery County, Ohio, 2004.

Higher numbers of overvotes occurred in precincts using this ordering of candidate names, which may have been due to the visual grouping effect the "Candidate Removed" line had on the remaining candidate options. Voters may have interpreted the two visual groupings as two separate races, one between Bush and Kerry, and the other between Peroutka and Badnarik. By selecting two candidates (in what a voter thought was two

separate races) for president, neither vote would actually have been recorded (Brennan Center for Justice, 2008). About 2.6% of the voters in the 33 Dayton precincts using this ballot rotation overvoted in the presidential race. In contrast, only 1.1% of voters in the Dayton precincts using other ballot rotations overvoted in that race (Jacobs, 2005).

The Brennan Center report (2008) also describes several cases where ballots failed to use shading and bolding to help voters differentiate between voting tasks. In Escambia County, Florida, the ballot used in the 2002 general election lacked any sort of shading to help voters differentiate between races. Residual vote rates with this unshaded ballot were higher than in neighboring Bay County, where the ballot's shading aided voters in differentiating between the various races (Figure 4).

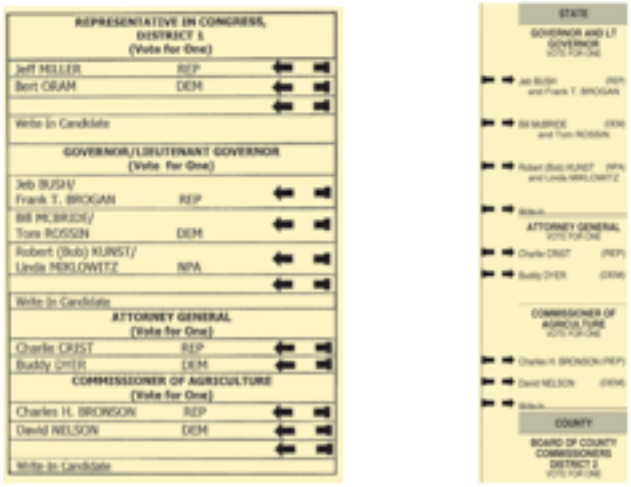


Figure 4. Ballot sections from Escambia County (left) and Bay County (right), Florida, 2002.

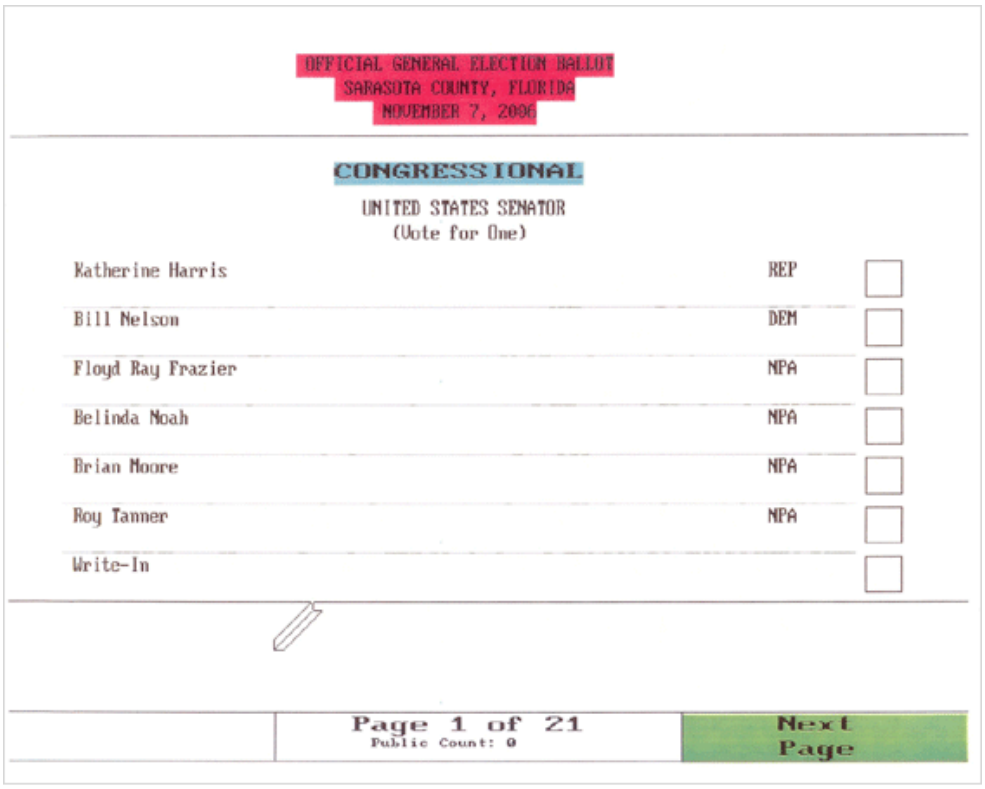
More specifically, the residual vote rate in Escambia County for the Attorney General race was 2.5%, whereas it was 2.2% in Bay County. Similarly, the residual vote rate in Escambia County for the Commissioner of Agriculture race was 4.6%, versus 4.3% in nearby Bay County (Brennan Center for Justice, 2008).

The influence of ballot formatting on the residual vote is not a new finding by any means. In fact, effects of inconsistent ballot headings in the 1976 general election in Los Angeles County, California, were much larger than any of the incidents described in preceding paragraphs. As can be seen in Figure 5, the 1976 Los Angeles County ballot used office title and instructions that were presented above candidate names for the presidential race, whereas that information was presented to the left of candidate names for the U.S. Senate race.

Figure 5. Section of the ballot used in Los Angeles County, California, 1976.

This inconsistent use of headings resulted in an exceedingly high 17.0% residual vote rate in Los Angeles County for the Senate race, whereas the rate statewide was only 4.1%. The number of votes lost in the county (436,864 votes) far exceeded the margin of victory for the Senate candidate in the state (246,111 votes) (Brennan Center for Justice, 2008).

Of all the incidents detailed in the Brennan Center report and described in preceding paragraphs, the 2000 "butterfly ballot" is perhaps the ballot design fiasco most well-known by the voting public. However, the actual difference in presidential residual vote rates between Palm Beach County and the state of Florida was quite small (3.3%) in comparison to similar effects seen in recent elections using poorly designed DREs. For example, a difference in residual vote rates of 11.4% was seen in the 2006 general election in Sarasota County, Florida. While the first DRE screen voters saw only had a single race on it, the second screen they saw had two races on it (Figure 6). The residual vote rate for US Representative, the top race on the two-race screen, was 13.9% for Sarasota County voters using DREs with this screen layout. In sharp contrast, the residual vote rate for this race was only 2.5% for Sarasota County voters using provisional or absentee ballots. The residual vote rate for this race in nearby Charlotte County was also 2.5%. This fact is quite important, because DREs in Charlotte County displayed the US Representative race alone on a separate page. In the Sarasota 2006 election, the number of votes lost far outnumbered the margin of victory: the margin of victory for US Representative was only 369 votes, whereas the number of residual votes was 18,413 (Brennan Center for Justice, 2008).



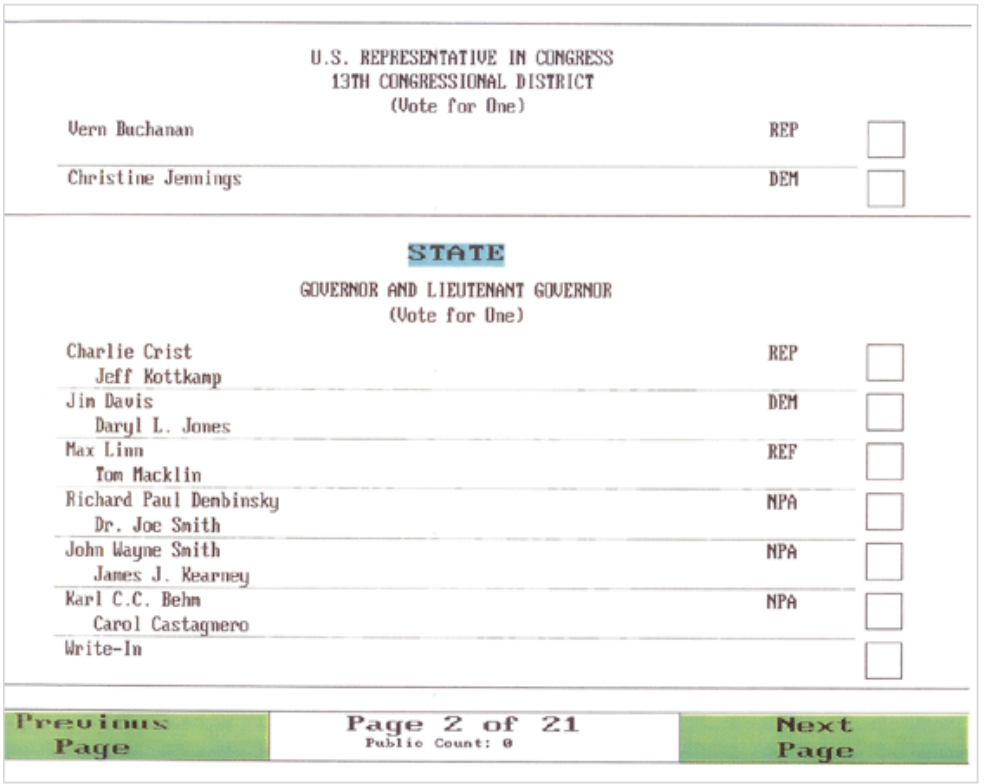


Figure 6. First two DRE screens of Sarasota County ballot, Florida, 2006.

An even larger effect of poor DRE ballot design was seen in the 2006 general election in Charlotte County, Florida, where the difference in residual vote rates for the Attorney General race between the county and the state was 16%. The DREs used in Charlotte County presented the races for Governor and Attorney General on the same screen (Figure 7).

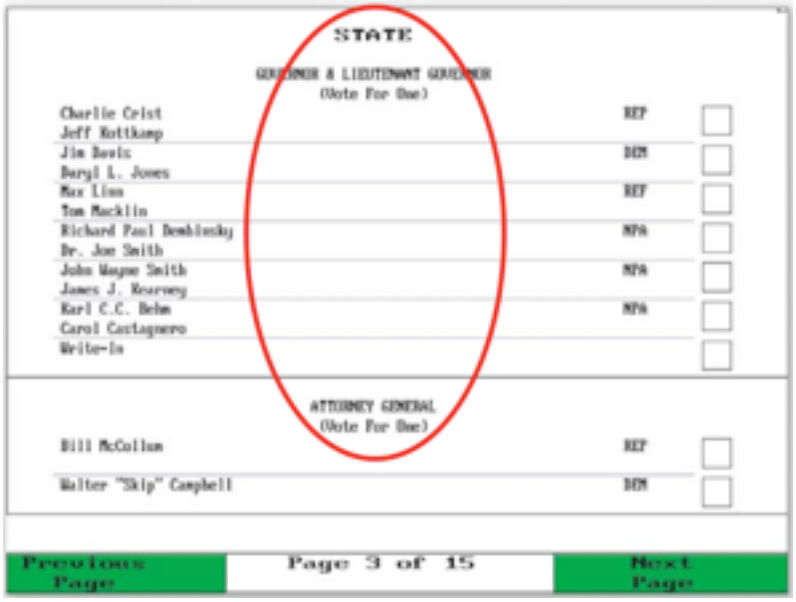


Figure 7. DRE screen from Charlotte County ballot, Florida, 2006.

The residual vote rate in Charlotte County for the bottom-most of those two races (i.e. Attorney General's office) was extremely high, 20.9%. That rate statewide was only 4.9%, and it was only 4.3% in nearby Sarasota County, where the ballot used displayed those two races on separate screens.

The Sarasota election gave rise to much controversy when competing explanations were proposed to account for the US Representative race results: malicious/

malfunctioning software, touchscreen miscalibration/insensitivity, voter apathy, and poor ballot design.

Although some have suggested that malicious/malfunctioning software could have been to blame in the Sarasota and Charlotte County incidents, it remains more than likely that ballot design had at least some effect on voter performance. Multiple hypotheses have been proposed regarding which specific design flaws were responsible, yet no studies have directly examined them. In addition to the desire to differentiate between these multiple hypotheses (which are discussed in the following section), the Sarasota County ballot was chosen as the basis for the current studies due to the sheer magnitude of its effects on election results, and also because it could be compared to an almost perfectly inverse design in the Charlotte County ballot. Finally, because DREs are more similar than older voting methods are to the graphical user interfaces (GUIs) traditionally studied in the field of HCI/HF, knowledge gained from the current studies has greater potential for wider applicability in the larger HCI/HF community.

Study rationale

Several different hypotheses have been proposed to explain the Sarasota County undervote rates seen with the critical two-race screen, yet there are insufficient data to support or refute any or all of them. Many people suggested that malicious or malfunctioning DRE software was to blame. These claims could not be proven, as DRE vendors did not disclose the necessary source code for examination by computer security experts; ES&S successfully invoked the trade-secret privilege during post-election

litigation (Amunson & Hirsch, 2008). Other people proposed that voters simply did not know anything about the candidates, and therefore intentionally omitted the race in question. This suggestion is unfounded, given the expensive and ubiquitous nature of the campaign for that office. At the time, it was the single most expensive non-presidential election in the history of the United States. Of greater interest and relevance for the current studies are the multiple hypotheses suggesting that various ballot design issues were at fault.

It has been hypothesized that "banner blindness" was to blame for the large number of voters who did not choose a candidate in the US Representative race in Sarasota County, 2006. Just as web searchers often fail to notice links at the top of web pages (Benway & Lane, 1998), so too could voters have failed to notice the race at the top of the DRE screen. This idea could be tested simply by displaying each race on its own screen. A second hypothesis proposed that inconsistent layout was at fault: the first screen voters saw only had a single race on it, therefore they were expecting only a single race on the second screen as well. This might be examined by varying the number of *races* on the first screen (one versus multiple), as well as by varying the number of *screens* before the critical two-race screen. A third hypothesis suggested that inconsistent use of highlighting was the problem: on the first screen, the race heading "Congressional" was highlighted in blue above the Senator race, but this heading was not present above the congressional race (US Representative) on the second screen. While the "Congressional" header was absent above the top race on this critical second screen, the "State" header above the bottom-most race was present and highlighted. Effects of

highlighting on visual attention could be examined in several ways. One could remove the highlighting and keep the race headers as they were initially, one could remove the highlighting while adding the missing Congressional header, or one could do away with the headers entirely.

Although descriptions of the possible explanations for the Sarasota County ballot effects abound, ranging from news articles (Doig, 2006) to columns by well-known usability experts (Nielson, 2007), no studies have directly examined the different hypotheses. Was it banner blindness, inconsistent layout, or inconsistent highlighting at play in Sarasota County? Was it an additive or interactive combination of the three? If so, which factor most influences human performance and why? Understanding the exact mechanisms by which particular design characteristics affect voter error remains an open empirical issue. While there is ample information detailing *what* happened in elections such as Sarasota County, there is no corresponding literature explaining exactly *why* voters made these errors. There is clearly a need for a theoretically sound, empirically generated explanation of *why* various types of poor ballot design consistently result in specific classes of voter errors. Rationale for the current studies stemmed from a desire to begin addressing this larger issue, by starting with a specific real-world ballot design problem. A brief summary is below, with a more detailed description in subsequent sections.

Study 1 sought to shed light on the viability of the various hypotheses discussed above. Participants voted on a DRE with a critical two-race screen similar to the 2006 Sarasota County ballot. Highlighting was manipulated, as well as the number of races

presented on the screen preceding the critical two-race screen. An ACT-R cognitive model was then constructed based on Study 1 human behavioral data. The model was then run with a novel variation of the Study 1 ballot, in an attempt to predict the human performance changes that one would expect to see with altered interface designs. Study 2 collected behavioral data to which model predictions were then compared.

STUDY 1

Method

Participants

Participants for Study 1 were 144 adults recruited from the larger Houston population. The primary recruiting method was a paid advertisement in the Houston Chronicle. Additional recruiting methods included placing a paid advertisement on Craig's List online, and contacting people who previously expressed interest in our research but had not yet participated. Participants were compensated $25 cash for the one-hour study. Participant eligibility requirements were as follows: must have normal or corrected-to-normal vision, be at least 18 years of age, be a US citizen fluent in English, and not have participated in any of our previous voting studies. Of the 144 total participants, 76 were female and 68 were male. Ages ranged from 18-84 years, with a mean age of 46.8 (*SD*=17.6 years), and median age of 48 years. Participants were diverse in terms of education, ethnicity, income, and voting experience.

Design

A mixed design with one within-subjects variable and three between-subjects variables was used. The within-subjects variable was "DRE versus punch card," since each participant completed the same ballot two times: once with a DRE and once with a punch card. While previous research tested three different non-DRE voting methods—paper ballots, lever machines, and punch cards—the current studies used only punch cards, as they most closely resemble the DRE in terms of task structure and visual presentation of information. Lever machines are full-face ballots: the entire ballot is displayed at once, and this information is ever-present. Similarly, voters see multiple races simultaneously when using bubble ballots (although these may not be entirely full-faced if the paper must be turned over). The simultaneous presentation of numerous races by a lever machine or bubble ballot would have been very different from the Sarasota County DRE ballot of interest, where only one or two races were presented simultaneously on the screens of primary interest. The maximum number of partisan races per screen was limited to three. By showing only one or two races per booklet page, punch cards easily mimicked the DRE's method of presenting only a limited amount of information at once. For this reason, punch cards were the sole non-DRE voting method used in Study 1. In addition to displaying the same number of races per page as the DRE, punch cards were also analogous to the DRE in terms of font styles, colors, spacing, etc. (Figure 8).

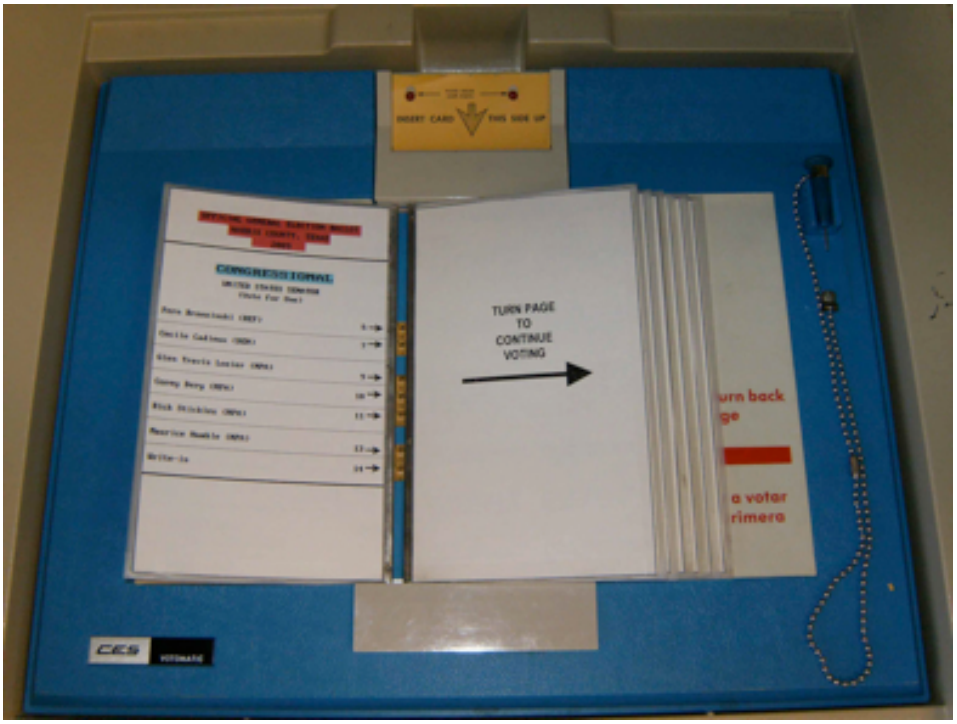


Figure 8. Picture of punch card used in Study 1.

The first between-subjects variable was information condition: a third of participants were in the *directed* information condition, while the remaining two-thirds were in the *undirected* information condition. In the *directed* condition, participants were given a sheet of paper listing exactly what selections they were to make while completing each ballot. Participants kept this paper with them while voting, and the experimenter stressed before each ballot that participants were to make only the choices listed, regardless of their personal political preference. More explicitly, because participants were instructed to intentionally omit certain races based on real-world roll-off rates (Nichols and Strizek, 1995), the directed condition in Study 1 could more accurately be termed "directed with roll-off." However, for the sake of simplicity, it will henceforth be referred to as the "directed" information condition.

In the *undirected* information condition, participants were offered a voter guide and allowed to make candidate selections based on personal political preference. Voter guides were similar to those used in previous studies, which were based on guides put forth by the League of Women Voters. All participants in the undirected information condition were instructed to vote consistently, i.e. to select the same candidates on their second ballot as they chose on their first ballot. Participants were reminded to vote consistently before receiving each ballot. The undirected information condition was then further subdivided: half of participants in the undirected condition were explicitly told to make a choice in *every* race. The other half of participants in the undirected condition were simply instructed to vote as they would in a real election. For example, "if you would normally make a selection in every race, please do so"; "if you normally skip races, please do not make selections for those races". Participants were randomly assigned to information conditions.

The second between-subjects variable was the number of races displayed on the first DRE voting screen, which was the page immediately following the instructions screen. Half of participants was two races on their first voting screen (Figure 9), while the other half of participants saw only a single race on their first voting screen (Figure 10).

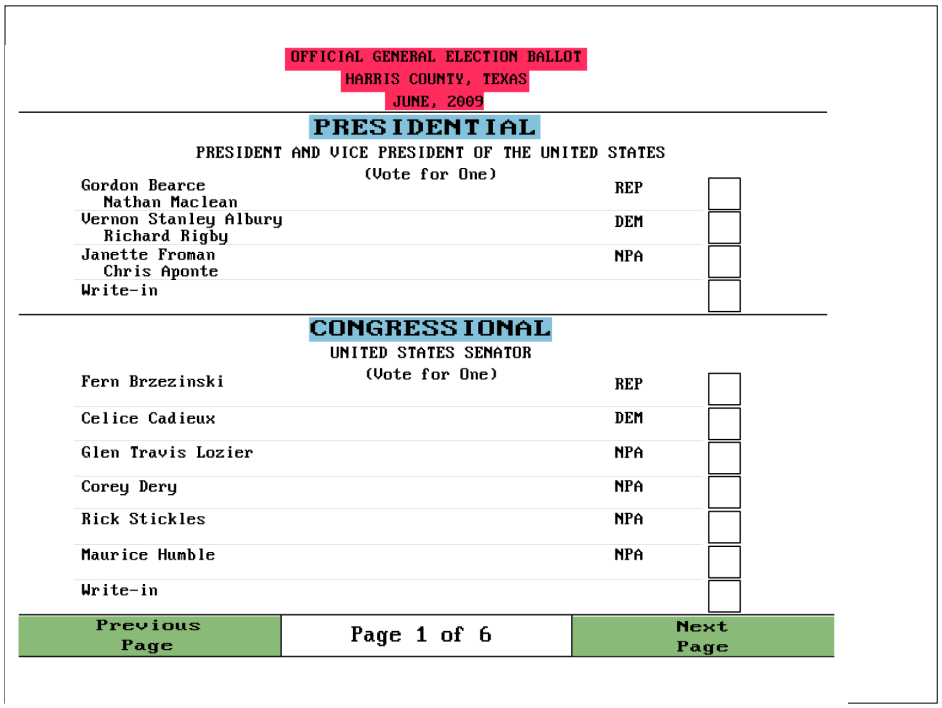


Figure 9. Two races on first screen of Study 1 DRE.

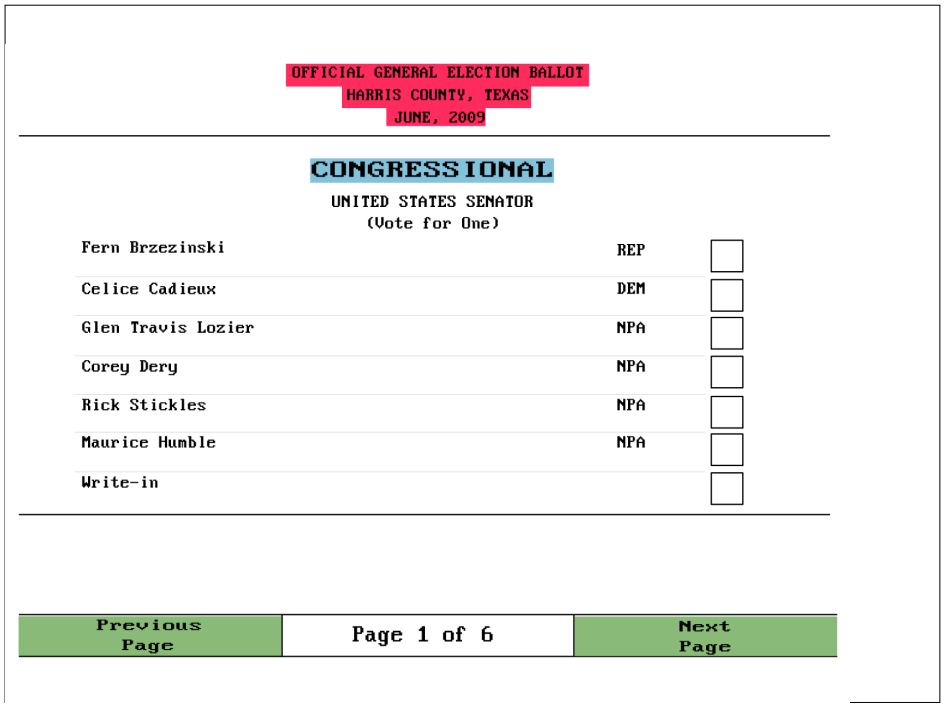


Figure 10. Single race on first screen of Study 1 DRE.

The final between-subjects variable was the manipulation of highlighting on the race category headings. Half of all participants saw the race headings of "Congressional" and "State" highlighted in blue as they were on the actual Sarasota County 2006 DRE ballot: this condition was called "original Sarasota highlighting". It should be emphasized that there was no heading at all above the US Representative race on the original Sarasota DRE ballot, nor was there in Study 1. In contrast to the original Sarasota highlighting, headings seen by the remaining half of participants were not highlighted at all, including both the race category headings and the overall ballot heading (see Appendices for screenshots of Study 1 DRE ballot in its entirety).

There were several dependent variables measured, which remain consistent with those from prior research: ballot completion time, errors, and satisfaction. These three variables are in keeping with the ISO usability metrics recommended by NIST: efficiency, effectiveness, and satisfaction. Overall ballot completion times were measured in seconds, and assessed the total time it took a participant to complete a single ballot. DRE ballot completion times were recorded automatically by the DRE software, whereas the experimenter recorded punch card ballot completion times by hand via stopwatch. In addition to overall DRE ballot completion times, *per-screen* and *per-race* times were measured for the DRE (in milliseconds). Of particular interest were per-screen times for the second voting screen, which presented the critical US Representative race.

The most important dependent variable in this study was error rates. The way an error was determined differed based on a participant's assigned information condition, since determining voter error hinges on knowing voter intent. In the directed information

condition, voters intend to make the selections listed on the sheet they are given. Therefore, any selections other than those listed are considered errors. If participants chose a different candidate than the one listed, that was termed a “wrong choice” error. If they failed to make a selection where one was indicated on their sheet, that was counted as an “undervote” error. Finally, if they made a selection for a race their sheet instructed them to omit, that was classified as an “extra vote” error.

In the undirected information condition, voters intend to choose the same candidates on the DRE and the punch card. Therefore, discrepancies between ballots were considered errors. Determining which voting method the error should be attributed to was accomplished via an exit interview. In the undirected information condition only, participants completed an exit interview after finishing their second ballot. An exit interview was necessary to determine voter intent by majority rule. For example, if a voter selected candidate A on the DRE, candidate B on the punch card, then stated that they voted for candidate A during the exit interview, the error would be attributed to the punch card. As previously described for the directed information condition, the same distinction was drawn between “wrong choice,” “undervote,” and “extra vote” errors in the undirected condition.

Regardless of information condition, there was a fourth type of error that voters could make with the punch card that could not occur with the DRE: an “overvote” error. An overvote occurred when a participant made more than the allowed number of choices for a race. For both Study 1 and Study 2, all races were single-selection races, meaning voters could choose at most one candidate per race. While DRE software prevents

selection of more than one candidate for the same race (lever machines prevent this error mechanically), punch cards (and paper ballots) do not have an analogous prevention mechanism. It is quite possible for participants to punch multiple holes (or fill-in multiple bubbles) for a single race, so Studies 1 and 2 both recorded "overvote" errors for the punch card and bubble ballot respectively, as well as the afore-mentioned "wrong choice," "undervote," and "extra vote" errors.

Two unique error types were of particular interest in Study 1. The first and most important error was whether participants skipped the critical US Representative race. This race was extremely important, as it was the top-most of the two races seen on the critical Sarasota-like race screen (Figure 11).

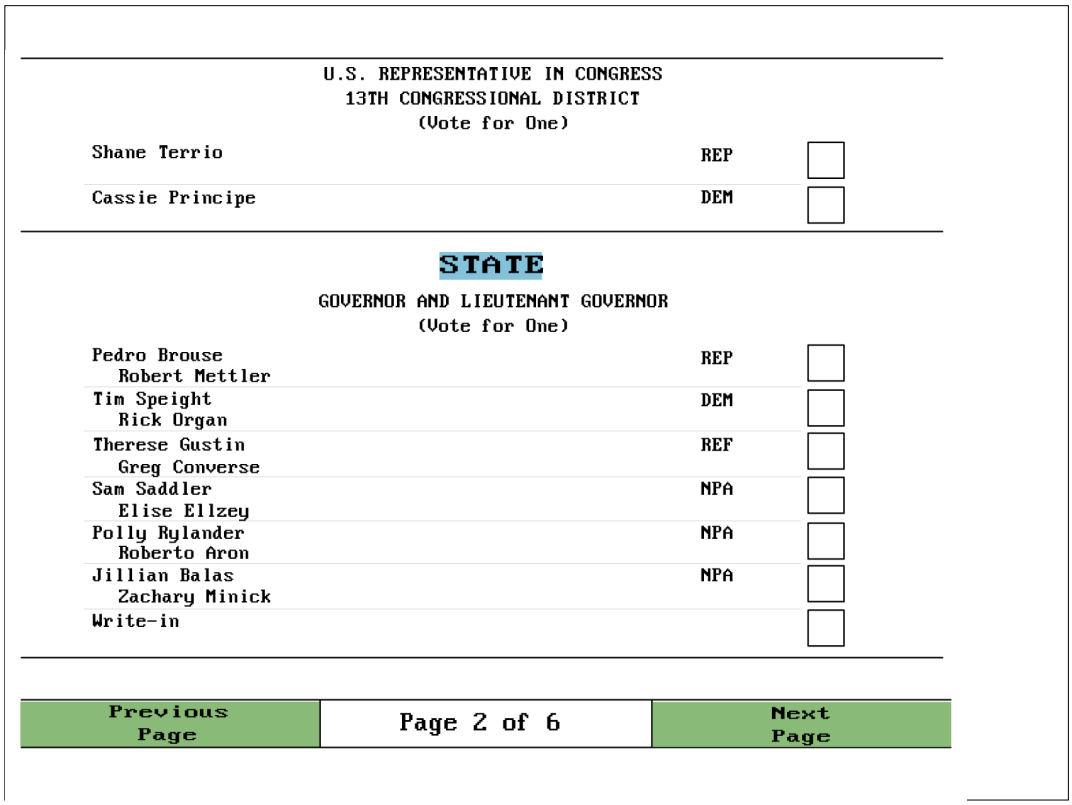


Figure 11. Study 1 DRE critical race screen.

The second error type of particular interest was whether voters failed to cast their ballot with the DRE. Failure to cast a ballot after making selections has been seen in real-world elections, called the "fleeing voter" problem, and has also been documented in recent experimental research (Greene, 2008).

In addition to measuring errors and ballot completion time, voter satisfaction was assessed. Satisfaction was evaluated using the System Usability Scale (Brooke, 1996). Participants completed two separate SUS questionnaires: the first was completed immediately after voting with the DRE, and the second immediately after voting with the punch card.

Materials and procedure

The punch cards and DRE hardware for the current studies were those used in previous research. Punch cards were purchased at auction from Brazoria County, Texas. The DRE was a desktop PC running Windows as its basic operating system, with a 17-inch LCD monitor. Adobe Flash was used to create and display the DRE ballots, and to record a time-stamped log of all user actions. A subset of the same fictitious candidate names from prior studies were used, although no propositions were included. Candidate names were arranged to make the DRE screens as similar as possible to the Sarasota 2006 ballot.

Participants began the study by completing the informed consent procedure, then were given a sheet of instructions to read. The experimenter then verbally reviewed the instructions and answered any participant questions. The exact wording of instructions

differed slightly between the two information conditions. As previously mentioned, participants in the directed information condition were instructed to vote by making only the selections listed on the sheet they were given. In the undirected condition, half of participants were instructed to make a selection in every race, while the other half were told to vote as they normally would in a real election. All participants in the undirected condition were reminded to be consistent in their selections across ballots.

After giving informed consent and receiving instructions, participants in the undirected information condition were offered a voter guide. If they chose to read it, they were given ample opportunity to do so, and were allowed to write on it and keep it with them as they voted. Instead of a voter guide, participants in the directed information condition received a sheet of paper listing all the selections they were to make while voting. After receiving the appropriate document for their assigned information condition, participants were shown to a voting station. DRE ballots and punch cards were both completed on waist-high tables; participants remained standing while voting in order to more closely simulate a real election.

After participants voted on the DRE, they completed a SUS questionnaire to assess their satisfaction with that voting method. The SUS is a simple 10-question scale, using questions like “I thought the system was easy to use” and “I felt very confident using the system.” People respond using a Likert scale of one to five to indicate whether they “strongly disagree” or “strongly agree” with each statement. Ratings are combined into a single number, ranging from 0 to 100, with higher numbers indicative of greater user satisfaction (Brooke, 1996). After completing the SUS for the DRE ballot,

participants then voted on the punch card, followed by another SUS questionnaire. Participants in the undirected information condition then completed an exit interview. During the exit interview, participants were given a sheet listing all the races and candidate names they saw while voting. Participants verbally indicated the selections they made and the experimenter circled those choices on her own sheet.

Regardless of information condition, all participants completed a final survey packet, which included voting-specific questions, as well as basic demographic questions. Examples include questions on number of elections voted in, experience with different voting methods, age, education, computer expertise, etc. Of note, participants were asked explicitly whether they voted in every race on the DRE, as well as whether they voted in every race on the punch card. After completion of the final survey, participants were debriefed and paid.

Results

Effectiveness

Of the 144 total participants run, seven were not included in the following error analyses. Five participants were not included due to problems recording their DRE data in Flash, while the remaining two participants were excluded due to being outliers, defined as having error rates greater than 15% on both voting methods. This outlier definition is in keeping with prior research.

Effectiveness: Critical Race Results

On the DRE, twenty-seven of 137 voters skipped the US Representative race their

first time on that screen, resulting in an initial omission rate of **19.71%**. Of those 27 voters, 20 corrected their critical race undervote error before casting their final ballot: 11 of those 20 voters corrected their mistake pre-review screen, while nine did so post-review screen. Seven of the 27 voters who skipped the US Representative race their first time on that screen did *not* correct their mistake, resulting in a final undervote rate of **5.1%**. Of note, three of those seven voters were in the *directed* information condition, meaning that even when they had a sheet of paper listing the US Representative race in front of them, they skipped it nonetheless. On the punch card, two participants undervoted in the critical race: one was in the directed information condition, while the other was in the undirected condition.

Effectiveness: What Predicted Skipping the Critical Race?

The only significant predictor of whether people initially skipped the critical race on the DRE was the number of races seen on screen one. Nineteen of the 63 voters (30%) who saw a single race on their first voting screen initially skipped the critical race. In contrast, only eight of the 74 voters (11%) who saw two races on their first screen did so. Critical race initial omissions were regressed on highlighting, number of races on screen one, information condition, education, and age (age as a continuous variable). All two-way interaction terms were tested initially in the first block of a logistic regression: none were significant. All main effects were then similarly evaluated: the model was not significant overall (Chi-square(7, $N = 137$) = 9.85, $p = .20$), accounting for 11% of the variance in predicting whether participants skipped the critical race. The main effect of number of Screen 1 races was significant, ($b = 1.17$, $p = .014$); people who saw a single

race on Screen 1 were more likely to skip the critical race than were people who saw two races on Screen 1. Although education did not reliably predict skipping the critical race (p = .48), voters with a postgraduate degree were somewhat less likely to make this error (Table 1). The main effect of age was not reliable (b = .001, p = .92), nor were the main effects of information condition (b = -.19, p =.71) or highlighting (b = .04, p =.93).

Table 1. Education levels for participants who initially skipped the critical race.

	High school or less	Some college	Bachelor's degree	Post-graduate degree
Count	8 of 26	9 of 46	8 of 47	2 of 18
Percentage	31%	20%	17%	11%

Effectiveness: Post-Completion Errors ("Fleeing Voter" Problem)

Thirteen of 137 voters (9.5%) failed to cast their vote with the DRE. Education significantly predicted the failure to cast a ballot: voters with less education were more likely to make this error (Table 2).

Table 2. Education levels for participants who failed to cast their DRE ballot.

	High school or less	Some college	Bachelor's degree	Post-graduate degree
Count	7 of 26	4 of 46	2 of 47	0 of 18
Percentage	27%	9%	4%	0%

Failed to cast error rates were regressed on highlighting, number of races on screen one, information condition, education, and age (age as a continuous variable). All two-way interaction terms were tested initially in the first block of a logistic regression: none were significant. All main effects were then similarly evaluated: the model was significant overall (Chi-square(7, $N = 137$) = 21.36, $p = .003$), accounting for 31% of the variance in predicting whether participants did or did not fail to cast their ballots. The fact that people with less education were significantly more likely to fail to cast their ballot ($p = .05$) is a result that would be expected given prior research. In contrast, finding that information condition significantly affected failure to cast a ballot ($b = 1.50$, $p = .03$) was rather unexpected: people in the directed condition were more likely to make this error, which has not been seen previously. Eight of the 42 people (19%) in the directed condition failed to cast their ballot; six of these people left it on the review screen and two of these people left it on the last race screen. Five of the 95 people (5%) in the undirected information condition failed to cast their ballots, four of whom left it on the review screen, whereas one person left it on the cast vote screen. However, the significant effect of information condition may have been an artifact of the experimental materials in

the directed condition, as the sheet listing the candidates people were to vote for in the directed condition instructed people to intentionally skip the last race. The main effect of age was not reliable (b = -.03, p = .17), nor were the main effects of highlighting (b = .49, p = .48) or number of Screen 1 races (b = -.74, p = .27).

Effectiveness: Overall Error Rates

Error rates were computed by dividing the number of errors made by the total number of opportunities for error. Participants who saw a single race on the first voting screen were presented with 13 races total, hence had 13 opportunities to err. Participants who saw two races on the first voting screen had a total of 14 opportunities to err. Error rates by voting method and error type are presented in Table 3 below.

Table 3. Error rates by voting method and error type.

	Extra vote	Undervote	Wrong Choice	Total Errors
DRE	0%	.59%	.75%	1.34%
Punch card	0%	1.97%	.74%	2.87%

Note that overvote error rates are not included in Table 3: while the punch card overvote rate was 0.16%, this type of error was not possible with the DRE. The punch card had a total error rate of more than twice that of the DRE (2.87% versus 1.34% respectively). A 2 (information condition) x 2 (header highlighting) x 2 (races on first screen) x 4

(education level) x 2 (age median split) x 3 (error type) x 2 (DRE vs. punch card) ANOVA found this main effect of voting method on error rates to be statistically reliable, $F(1, 83) = 14.62$, $p < .001$.

Of greater interest was the interaction between voting method and error type: while wrong choice and extra vote errors were comparable between the two voting methods, the punch card undervote error rate was much higher than was the DRE's. This interaction between voting method and error type was reliable, $F(2, 166) = 14.47$, $p < .001$ (Figure 12).

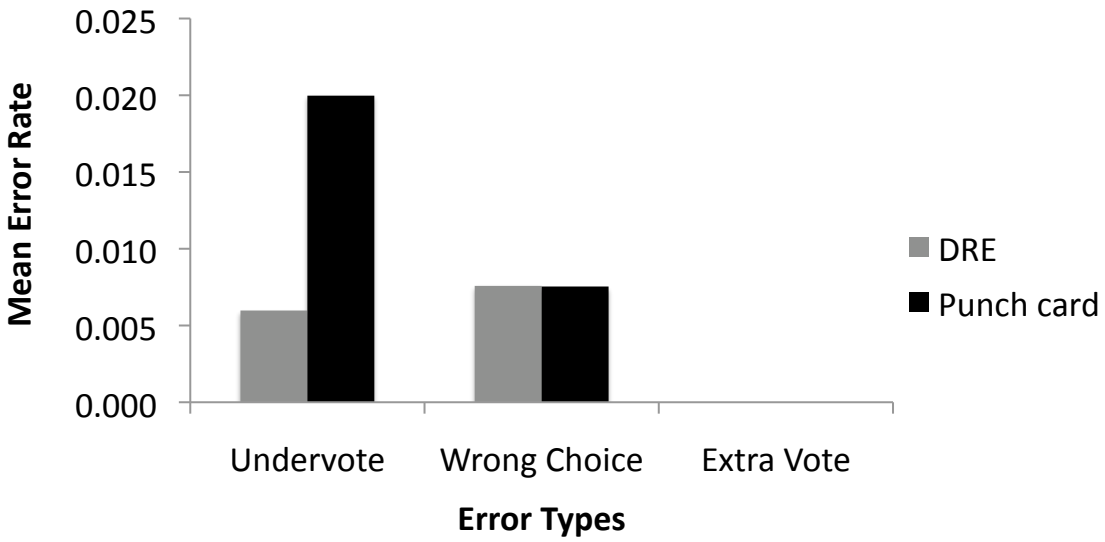


Figure 12. Interaction between voting method and error type on total error rates.

Total error rates between voting methods were similar in the absence of

highlighting. When header highlighting was present however, punch card error rates increased; DRE error rates were not similarly affected by highlighting (and in fact decreased slightly when highlighting was present). This interaction between voting method and header highlighting on error rates was reliable, $F(1, 83) = 18.86$, $p < .001$ (Figure 13).

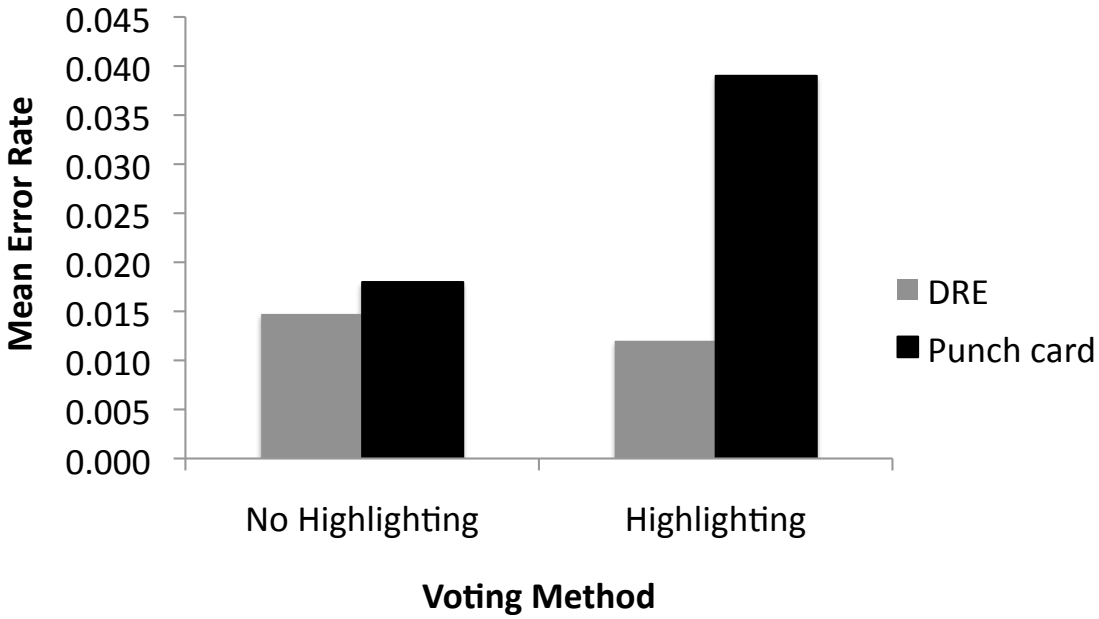


Figure 13. Interaction of voting method and highlighting on total error rates.

Voting method error rates were also differentially affected by education level: while DRE error rates were approximately the same across education levels, punch card error rates were greater for those voters with a high school education or less (Figure 14). This interaction between voting method and education level on error rates was reliable, $F(3, 83) = 12.31$, $p < .001$, and was decomposed with a post-hoc custom interaction contrast $F(1, 127) = 8.98$, $p = .003$.

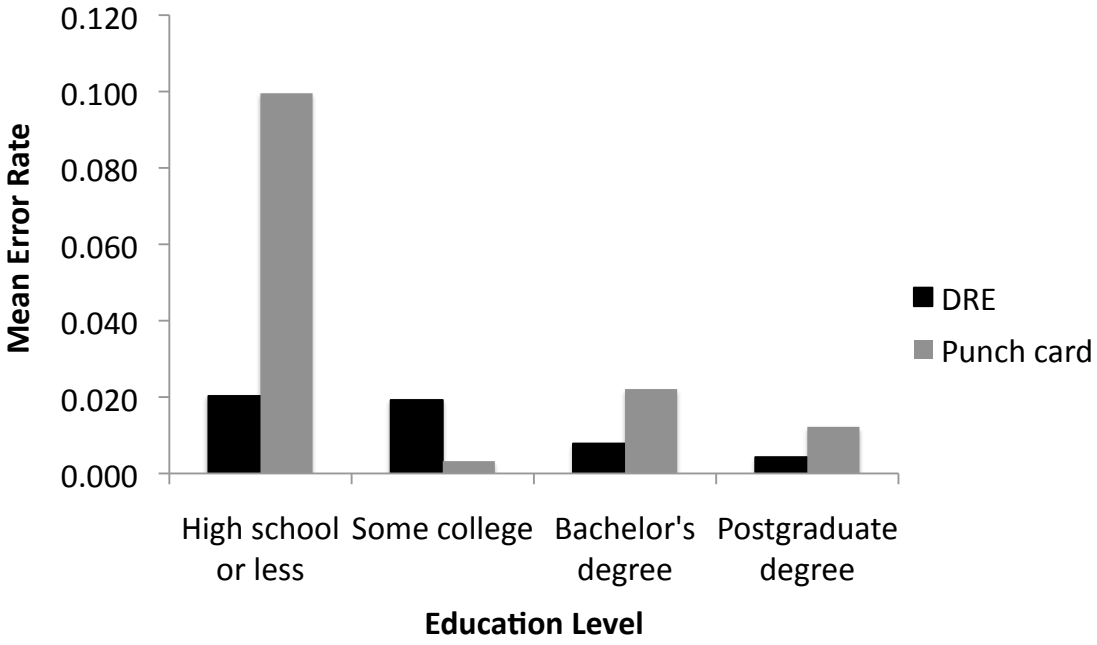


Figure 14. Interaction of voting method and education level on total error rates.

In addition to increased punch card error rates, voters with a high school or less education level also suffered from an increase in undervote error rates (Figure 15). This interaction between education level and error type was reliable, $F(6, 166) = 13.08$, $p < .001$, and was decomposed with a post-hoc custom interaction contrast $F(1, 127) = 17.43$, $p < .001$.

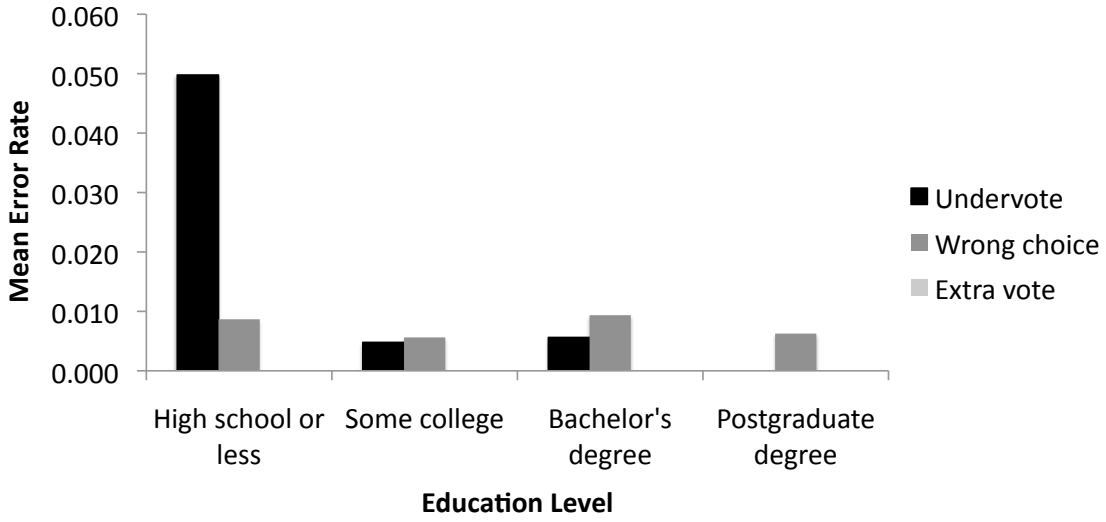


Figure 15. Interaction of education level and error type on total error rates.

While the presence of header highlighting resulted in increased total error rates for voters with a high school or less level of education, highlighting did not similarly affect error rates for the other education levels (Figure 16). This interaction between education level and header highlighting was reliable, $F(3, 83) = 3.47$, $p = .02$.

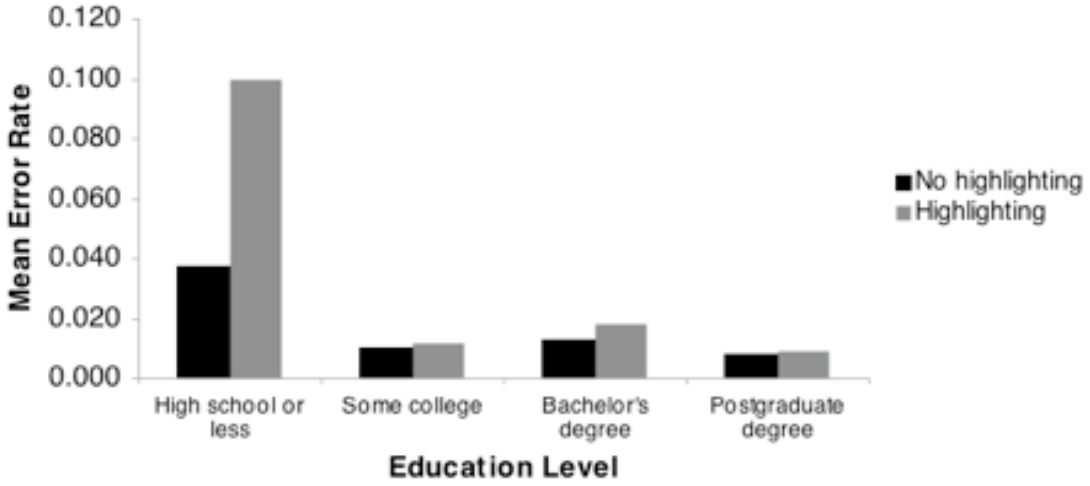


Figure 16. Interaction of education level and header highlighting on total error rates.

Given the nature of the previously described interactions involving education, it was not surprising that education level alone significantly impacted total error rates, with the high school or less education group having an error rate significantly greater than the mean error rate of the other three education levels combined (Figure 17). The main effect of education level on total error rates was reliable, $F(3, 83) = 9.19$, $p < .001$.

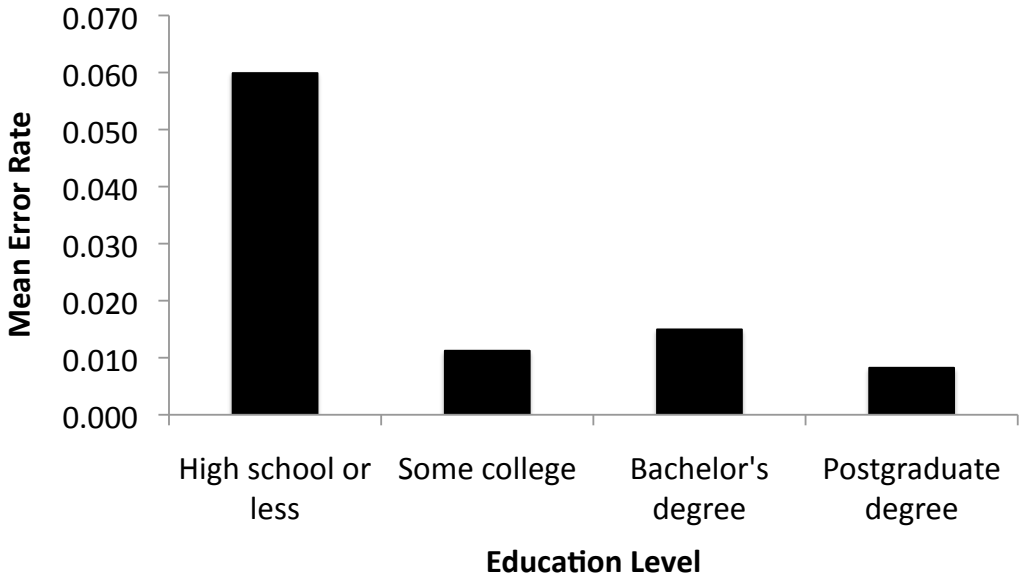


Figure 17. Main effect of education level on total error rates.

Total error rates were also significantly impacted by header highlighting alone: error rates increased reliably when highlighting was present, $F(1, 83) = 6.79$, $p = .01$ (Figure 18).

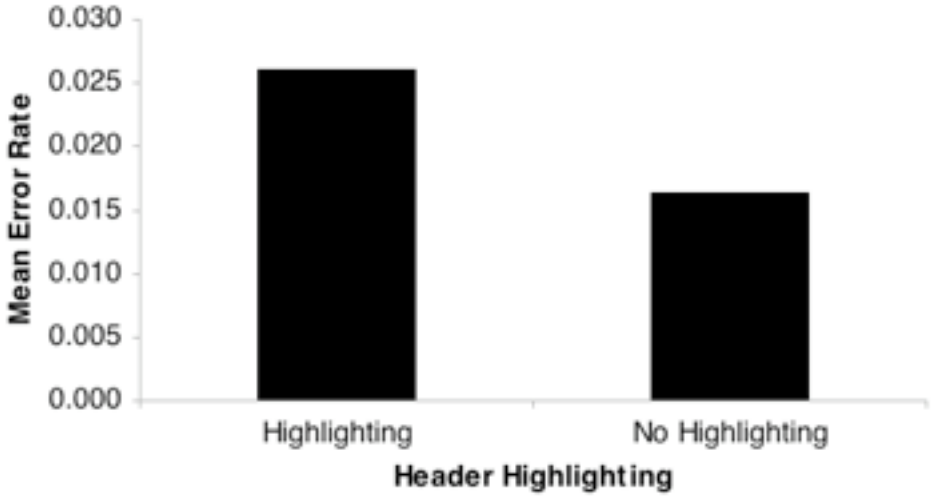


Figure 18. Main effect of header highlighting on total error rates.

Of even greater importance than the main effect of highlighting on total error rates was its interaction with error type: while wrong choice and extra vote error rates were similar regardless of highlighting condition, undervote error rates were noticeably greater with highlighting than without (Figure 19). This interaction between highlighting and error type was reliable, $F(2, 166) = 10.08$, $p < .001$.

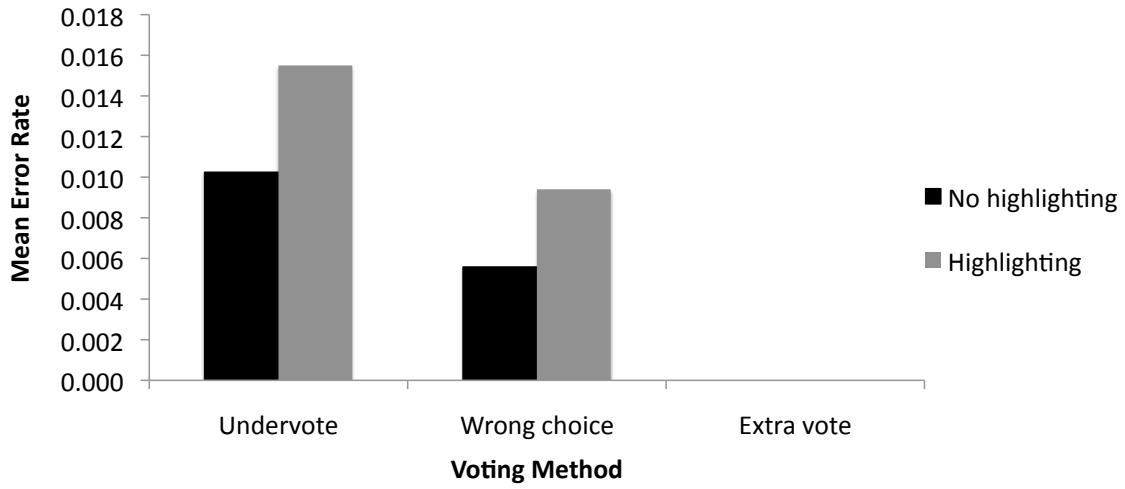


Figure 19. Interaction of error type and header highlighting on total error rates.

Finally, undervote error rates were significantly greater than wrong choice error rates, which in turn were greater than extra vote errors (Figure 20). This main effect of error type was reliable, $F(2, 166) = 17.14$, $p < .001$.

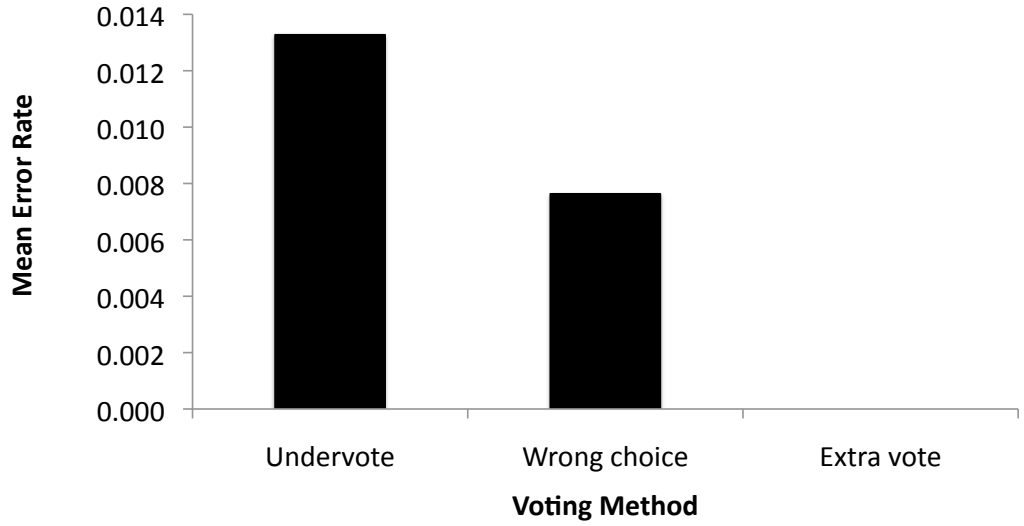


Figure 20. Main effect of error type on total error rates.

Intentional Undervotes

Not all undervotes are necessarily errors: a crucial distinction must be drawn between undervote *errors* versus *intentional* undervotes. In real elections, the two cannot be distinguished from one another, as voter intent is unknown. However, the design of this laboratory study made such a distinction possible. Since participants in the directed information condition were told explicitly which candidates to vote for (and which races to skip), intentional undervote rates were examined for those participants in the undirected information condition only. Ninety-six participants were thus included in the following analyses. The intentional undervote rate ranged from 0 to .692, with a mean of .047 (*SD* = .118) (Figure 21).

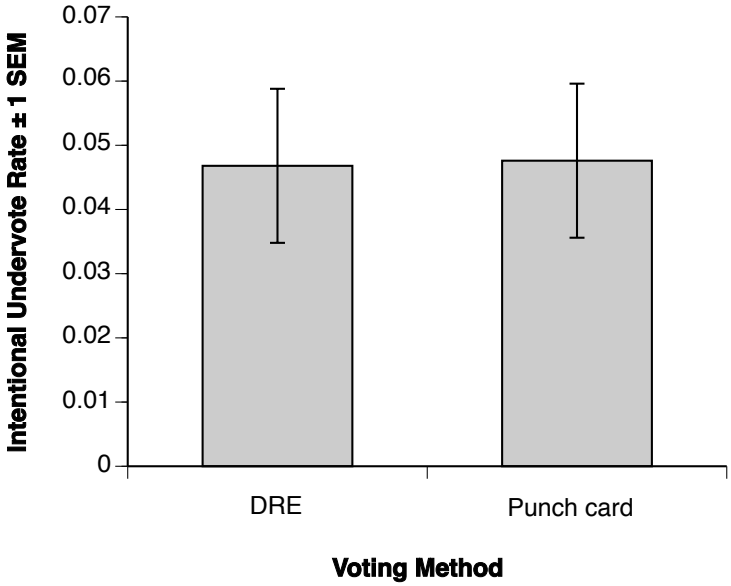


Figure 21. Overall mean intentional undervote rates.

Intentional undervotes for races seven and nine (State Representative and Charter Review Board District 2) were significantly higher than the mean intentional undervote rate for all other races on the ballot . These were the only two races on the ballot that did not have a Democratic candidate. The State Representative race offered one REP candidate and a Write-In option, whereas the Charter Review Board District 2 race offered one REP candidate and one NPA candidate. During the exit interviews, if participants said they skipped a race, the experimenter asked a scripted follow-up question to determine why they chose to skip the race in question. For the two races in question, participants consistently stated they did not make a selection because no Democratic candidate was listed. This main effect of race number was reliable, $F(2.50, 167.68) = 12.21$, $p < .001$, as seen with a 2 (header highlighting) x 2 (races on first screen) x 4 (education level) x 2 (age median split) x 13 (race number) x 2 (DRE vs. punch card) ANOVA. Note that because intentional omission rates for the first two races

(President/Vice President and Senator) did not differ significantly—in fact, they were both zero—the two were averaged together. Otherwise, depending on whether participants saw one versus two races on their first voting screen, the number of levels for the within-subjects variable "race number" would have been 14 for half of all participants, and 13 for the others.

Residual Vote Versus True Error Rates

One unique benefit of this type of laboratory study is the opportunity it provides to *directly* measure effectiveness by recording actual error rates. In contrast, during real elections, effectiveness is assessed only *indirectly* via the residual vote. The residual vote is an indirect measure of accuracy commonly used in Political Science. It reflects the number of "lost" or unrecorded votes, and is computed as the difference between voter turnout and the number of valid votes cast for any given race. It is comprised of overvote errors, undervote errors, and intentional undervotes. Since voter intent cannot readily be determined in real-world elections due to privacy concerns, the residual vote does not address wrong choice errors, nor does it distinguish undervote errors from intentional abstentions.

The true error rate for this study combined overvotes, undervotes, and wrong choice errors (recall that there were no extra vote errors, else they too would have been included). To calculate what would have been reported as the residual vote rate, overvotes, undervotes, and intentional undervotes were summed. Only the undirected information condition was of interest for these analyses, as it was most representative of a

real election scenario. A distinction is often drawn between the residual vote rate for the top-most race on a ballot versus the remaining "down-ballot" races. Although previously mentioned, it should be emphasized that residual vote rates of greater than 1% for top-of-the-ticket races are very rare, especially for President.

For those participants who saw only a single race on their first voting screen, the true error rate for the top-most race was .047 (*SD* = .183), whereas the residual vote rate would have been reported as .012 (*SD* = .076) in a real election (Figure 22). The two measures of accuracy were not reliably different from one another, $t(42) = 1.78$, $p = .08$; the correlation between the two measures was reliable, $r(42) = .81$, $p < .001$.

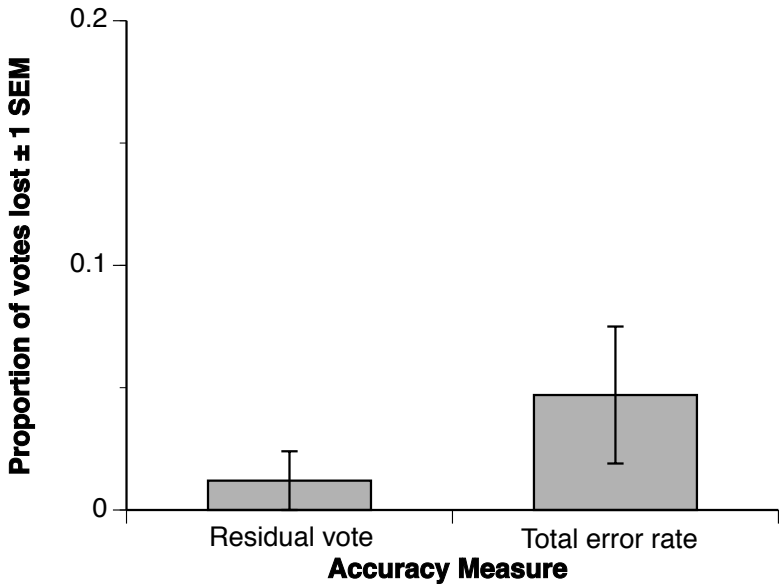


Figure 22. Comparison of direct and indirect accuracy measures for top-most race only, when participants saw a single race on first voting screen.

For those participants who saw two races on their first voting screen, the true error rate for the top-most race was .029 (*SD* = .118), whereas the residual vote rate would have been reported as .010 (*SD* = .069) in a real election (Figure 23). The two measures of accuracy were not reliably different from one another, $t(51) = 1.43$, $p = .20$; the correlation between the two measures was reliable (albeit lower than was the correlation for those voters who saw only a single race on the first screen), $r(51) = .57$, $p < .001$.

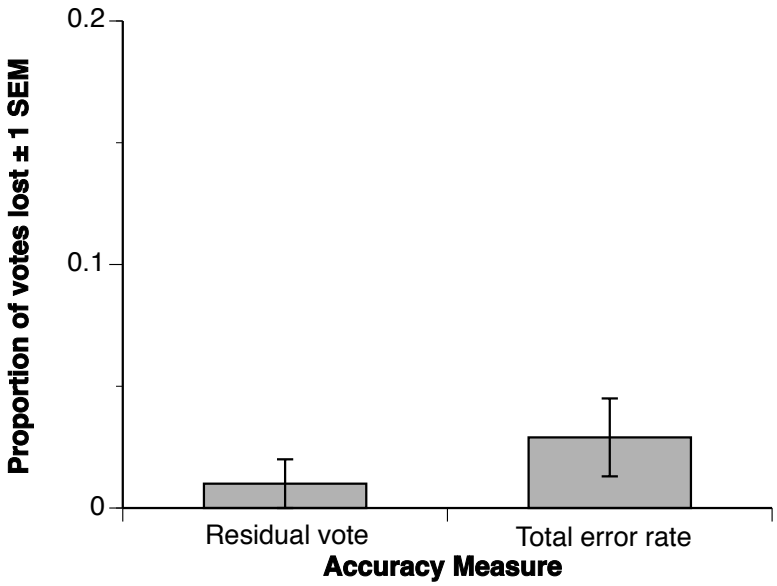


Figure 23. Comparison of direct and indirect accuracy measures for top-most race only, when participants saw two races on first voting screen.

When examining the remaining down-ballot races combined, for those participants who saw only a single race on their first voting screen, the true down-ballot error rate was .024 (*SD* = .067), whereas the residual vote rate was significantly higher, at .090 (*SD* = .157) (Figure 24). The two measures of accuracy were significantly different from one another, $t(42) = 2.68$, $p = .01$. Furthermore, the correlation between the two was low and unreliable, $r(42) = .15$, $p < .001$.

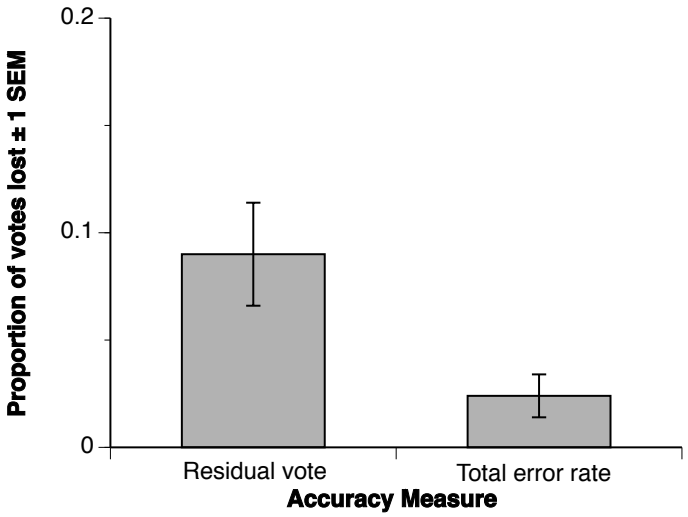


Figure 24. Comparison of direct and indirect accuracy measures for down-ballot races combined, when participants saw a single race on first voting screen.

For those participants who saw two races on their first voting screen, the true error rate for the remaining down-ballot races combined was .016 (*SD* = .033), whereas the residual vote rate was .040 (*SD* = .096) (Figure 25). The two measures of accuracy were not reliably different from one another, $t(51) = 1.71$, $p = .09$. However, the correlation between the two was quite low and unreliable, $r(51) = .07$, $p = .64$.

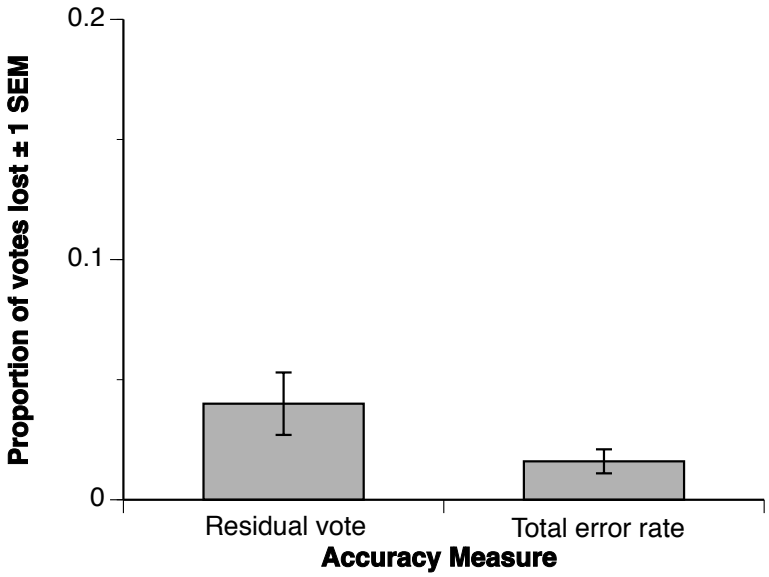


Figure 25. Comparison of direct and indirect accuracy measures for down-ballot races combined, when participants saw two races on first voting screen.

Efficiency

Data from 11 participants were not included in the following efficiency analyses: five participants were not included due to problems recording their DRE data in Flash, while the remaining six participants were excluded due to being outliers. Timing outliers were defined as observations falling outside three interquartile ranges (IQRs) from the 25th or 75th percentiles of the distribution of ballot completion times, a definition consistent with that used in prior research. Overall ballot completion times were similar between voting methods. The mean ballot completion time for the DRE was 194 seconds (*SD* = 128 seconds), while the mean ballot completion time for the punch card was 185 seconds (*SD* = 93 seconds) (Figure 26).

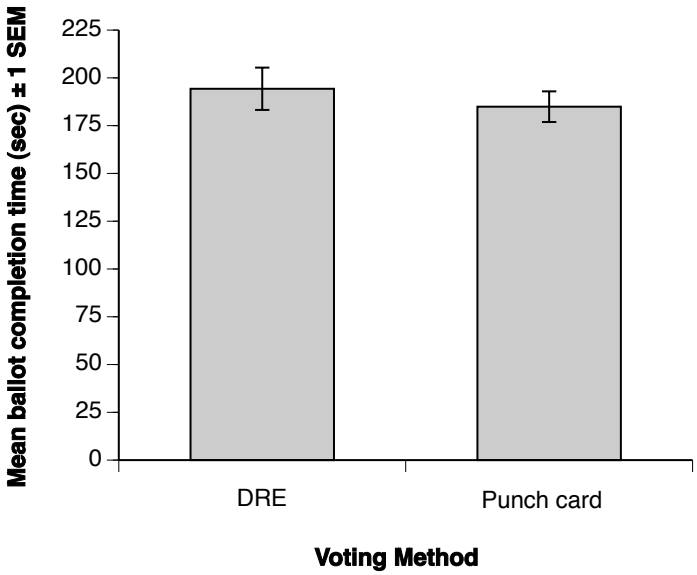


Figure 26. Overall mean ballot completion times.

Although voting method did not significantly impact ballot completion times, education level and information condition together did have an effect on completion times: in the undirected information condition, as education level increased, so too did ballot completion times (Figure 27). This effect was due to the amount of time that more highly educated voters spent reading the voter guide and referring back to it while voting. In contrast, voters with less education (especially those in the high school or less category) tended not to read the guide and instead did more straight-party ticket voting (mostly Democratic). This interaction of education level and information condition on ballot completion times was reliable, $F(3, 79) = 3.33$, $p = .02$, as found with a 2 (information condition) x 2 (header highlighting) x 2 (races on first screen) x 4 (education level) x 2 (age median split) x 2 (DRE vs. punch card) ANOVA.

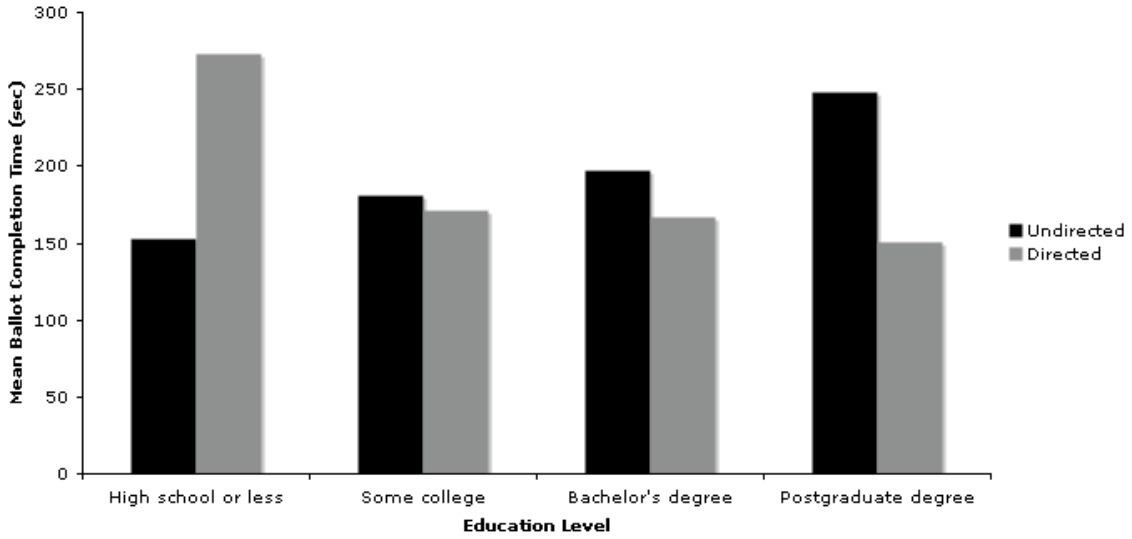


Figure 27. Interaction of education level and information condition on ballot completion times.

In addition to examining overall ballot completion times, one can also look at per-race times for the DRE. Of greatest interest are times for those races on the first and second screens only (Figure 28).

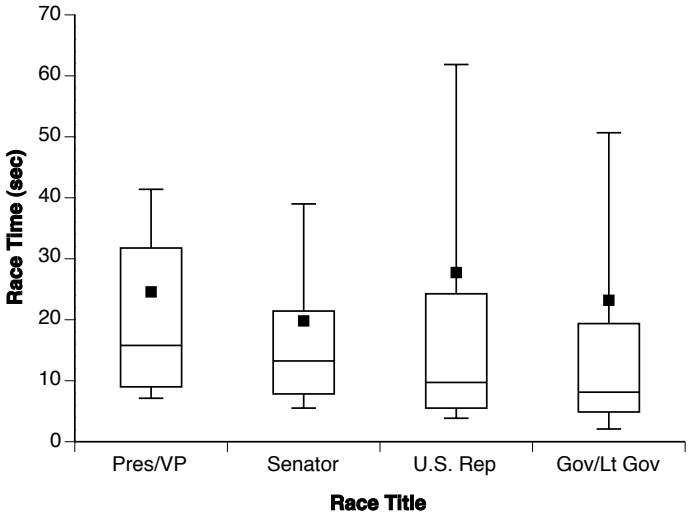


Figure 28. Per-race times for races on the first two DRE screens only (Pres/VP and Senator on screen one; US Rep and Gov/Lt Gov on screen two).

Recall that half of participants saw two races on their first voting screen (both the Presidential and Senatorial races), whereas the other half of participants saw only a single race (Senatorial). Per-race times were not significantly different between the Presidential and Senatorial races for those participants who saw two races on screen one. They were therefore averaged together to create a single variable representing the mean per-race time those participants spent on the first voting screen. This allowed for more direct comparison with those participants who saw only a single race on screen one, since levels of the within-subjects variable "race number" were then equivalent across conditions. (No other per-race times were combined, since all participants saw the remaining races on the ballot, regardless of their exact experimental condition.) Per-race times for the second voting screen were of even greater interest, as this screen presented both the critical race (US Representative) and the race for Governor/Lt. Governor.

Per-race times for the first two voting screens (i.e. the newly created "mean screen one" timing variable, US Representative race time, and Governor/Lt. Governor race time) were then analyzed in a 2 (information condition) x 2 (header highlighting) x 2 (races on first screen) x 4 (education level) x 2 (age median split) x 3 (race number) ANOVA. Per-race times for those races on screens one and two were significantly longer in the undirected information condition than in the directed condition, $F(1, 72) = 6.10$, $p = .02$. Although per-race times for screens one and two were somewhat slower with original Sarasota highlighting than without header highlighting, this difference was not statistically reliable at the conventional .05 alpha level, $F(1, 72) = 3.83$, $p = .06$.

Satisfaction

DRE SUS data from all 144 participants were included in the following analyses. However, one participant's punch card SUS data were not used, as the participant failed to complete the survey in its entirety. Overall, SUS ratings for the DRE were nearly thirty points higher than punch card SUS ratings (Figure 29). Whereas the mean DRE SUS score was 93.3 (*SD* = 10.1), the mean punch card SUS score was only 64.9 (*SD* = 21.8). Furthermore, participants were much more variable in their opinion of the punch card. In addition to having a standard deviation more than twice that of the DRE, the range of SUS scores for the punch card was much wider. Whereas punch card SUS ratings had an abysmally low minimum value of 5 (max of 100), DRE ratings ranged from 50 to 100.

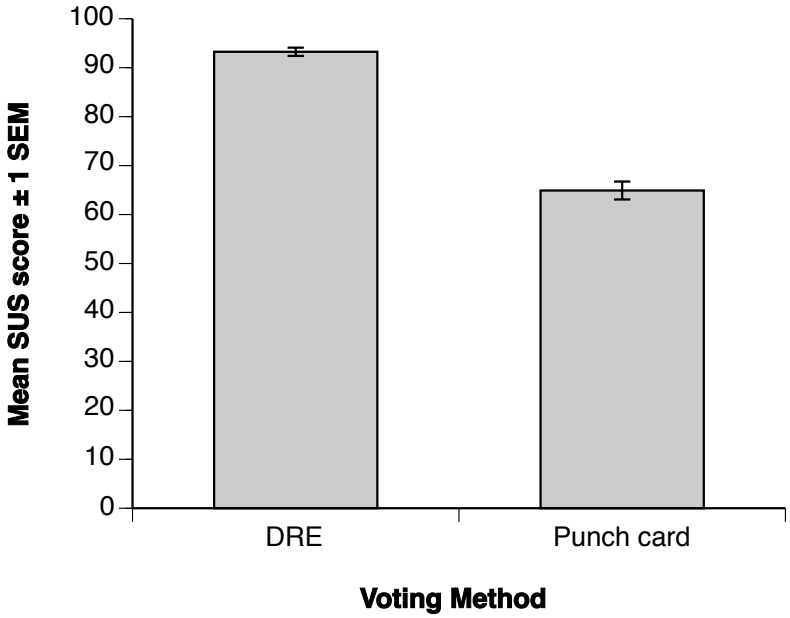


Figure 29. Main effect of voting method on satisfaction.

The main effect of voting method on SUS scores was reliable, $F(1, 89) = 157.92$, $p < .001$, as seen with a 2 (information condition) x 2 (header highlighting) x 2 (races on first screen) x 4 (education level) x 2 (age median split) x 2 (DRE vs. punch card) ANOVA. Although age had no effect on DRE SUS scores, it significantly influenced punch card ratings: younger participants (those below the median age of 48 years) rated the punch card reliably lower than did older participants (above the median age of 48 years) (Figure 30). This interaction of voting method and age on SUS scores was reliable, $F(1, 89) = 5.01$, $p = .03$.

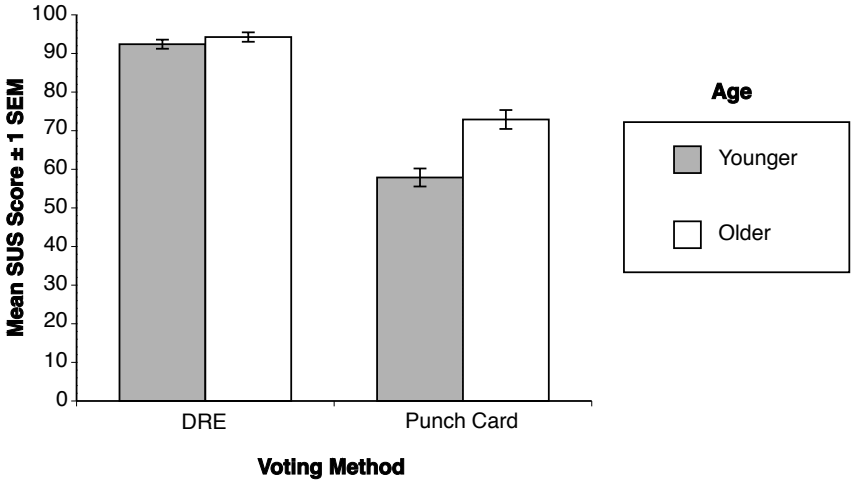


Figure 30. Interaction of voting method and age (median split) on satisfaction.

Discussion

While DREs offered no objective usability benefit in terms of efficiency, DREs received significantly higher subjective usability ratings than did punch cards. This voter preference for the DRE is not surprising given prior research. Also seen in prior research (Greene, 2008), was the significance of education in predicting the failure to cast a ballot; participants with lower levels of education (mainly those with high school or less) were more likely to make this specific type of post-completion error. Total error rates were reliably greater for the punch card than the DRE; an analogous relationship was found with undervote error rates. These effects may be due in large part to the difficulty some voters had inserting the punch card ballot properly. If not inserted all the way into the punch card holder such that the red tabs hold it in place, the ballot holes to be punched will not align properly with the associated candidate names. It should be emphasized that punch card error rates in Study 1 are a "best case" scenario. These punch cards were hand-scored; choices were counted as correct as long as voter intent could be inferred. Participants were therefore given credit for choices that machines may not have registered correctly due to pregnant, hanging, or dimpled chads. It is likely that punch card error rates may actually be worse in real elections.

Furthermore, punch card error rates were reliably higher when header highlighting was used, whereas DREs were not affected. This may be due in part to the size of the punch card ballot versus the DRE screen. When highlighting was present, the punch card ballot appeared more visually "busy" than did the DRE. Additionally, highlighting increased total error rates for those voters with a high school or less

education; punch card error rates were also reliably greater for those voters, whereas DRE error rates were fairly consistent across education levels.

When comparing true error rates with what would have been reported as the residual vote in a real election, significant differences were seen between the two measures of accuracy for down-ballot races. It is encouraging that the two measures did not differ significantly for the top-most race on the ballot, yet the fact that correlations between the residual vote rate and true error rates were low and unreliable for down-ballot races should be emphasized. These non-significant correlations suggest that the residual vote may not be as tightly coupled to true error rates as has been typically assumed. This in no way is meant to suggest that the residual vote does not provide useful information—in fact, it is difficult to imagine a viable means of replacing the residual vote with a more direct measure of accuracy without compromising voters' privacy.

The single most important result from Study 1 were the number of undervote errors voters made on the critical US Representative race. The critical race results seen in Study 1 offer strong support for the claim that human factors played a significant role in the Sarasota County 2006 election results. The initial omission rate for the US Representative race was 19.71%, while the final undervote rate was 5.1%. While the highlighting of race headers did not reliably predict initial omissions of the critical race, the number of races presented on the first voting screen did: when voters saw two races on the first screen, they were less likely to omit the critical race on the following screen than were voters who saw only a single race on the first screen. The critical race initial

omission rate of 19.71% in Study 1 was greater than the 13.9% residual vote rate for that race during the actual election in Sarasota County. However, the final undervote rate of 5.1% in Study 1 was less, a disparity which clearly warrants discussion.

The disparity between the final critical race undervote rate found in the laboratory versus that seen in the election is most likely attributable to the delayed touchscreen response on the Sarasota County DREs. As previously mentioned, the ES&S iVotronic systems used in Sarasota County suffered from a known problem with a "delayed response to touch." This was a problem of which state and county election officials were aware even before election day: ES&S stated in an August 15, 2006 letter to Florida election officials that the company had found an issue with the "smoothing filter" on the iVotronic touchscreens, which caused the machines to exhibit a "delayed response to touch." The necessary source code update and state-level certification needed were not completed, despite the manufacturer's statement that it planned to finish said repairs before the November 2006 General Election. Furthermore, Sarasota County's own records (incident reports kept by the Supervisor of Elections) described issues with the iVotronics on Election Day that necessitated multiple machines being "taken out of service on Election Day," issues such as their slow response to touch and need for a hard or extended touch before recognizing a vote. Finally, a machine-by-machine statistical analysis of election returns by the MIT political-science department chair, Charles Stewart II, found the undervote problem to be worst on touchscreens calibrated and set up on days the election staff was most busy (Amunson & Hirsch, 2008).

The letter from ES&S, Sarasota County's records, and Stewart's analyses all support the claim that the iVotronics in question suffered from touchscreen insensitivity/ miscalibration. However, the reason why such touchscreen issues would account for the increased undervote rate in Sarasota County relative to Study 1 requires further explanation. A recent study by Mascher, Cotton, & Jones (2010) directly addresses this, finding that a delayed touchscreen response decreased the number of voter navigation events. The finding that touchscreen insensitivity deterred proofreading was an unexpected result, as the primary aim of the study was to identify the voter interaction data that could be collected to enable informative investigation of touchscreen malfunction while simultaneously protecting voter privacy. In order for DRE event logs to actually be useful in examining interface/machine issues post-election, they must be sufficiently detailed. However, if event logs are so detailed they make it possible to reconstruct how someone voted, the right to a secret ballot is compromised.

To balance these competing requirements, Mascher et al. (2010) recorded three types of data: button types, relative locations, and timestamps. Two type of locations that a voter touched were recorded: button touches and background touches. For background touches, the time of touch was recorded, whereas the location was not. For button touches, the type of button but not the identity was recorded, as were relative rather than absolute touch coordinates. The authors examined compressed ballots, dishonest summary screens, touchscreen miscalibration, and delayed touchscreen response. Obviously the latter manipulation is of primary interest for purposes of the current discussion. When manipulating touchscreen delay, the authors added a delay of 100 or

250 ms to the display-feedback time to simulate different degrees of touchscreen insensitivity. The addition of a 250 ms delay resulted in a decrease in the number of extra navigation events by subjects in that group, as compared to the 100 ms delay and control groups, which were similar to one another. Markedly fewer participants in the 250 ms delay group reviewed contests than did participants in the 100 ms delay or control groups. The authors offered two possible explanations for the decrease in contest review behavior. Either the increased delay makes voters more confident that they do not need to review their contest choices, or the delay is so annoying that voters are dissuaded from reviewing their choices: given that the authors received several vehement complaints from voters in the 250 ms delay group, they hypothesized that the latter explanation was more likely.

One cannot help but agree, as a delay is extremely frustrating to users of any system—simply imagine a web session using dial-up rather than DSL or cable. An increased delay in system response time results in a large cost associated with navigation events by the user. In the case of the Sarasota 2006 election, a delayed touchscreen response likely reduced the number of voters who would take the additional time and endure the additional frustration associated with paging backwards to correct an undervote (whether or not voters noticed an undervote on the review screen is a separate issue altogether, see Everett, 2007).

In conjunction with other research (Mascher et al., 2010), Study 1 results strongly suggests that the 2006 Sarasota election was almost certainly a human factors problem. Nearly four years after the election, there are finally data that directly address the

question of whether poor ballot design affected Sarasota election results. Although the Jennings versus Buchanan court case lasted nearly 15 months, the issue of ballot design was never examined. Study 1 directly addressed this practical real-world issue, while also setting the stage for the introduction of cognitive modeling as a technique to potentially offer a more theoretical examination of the cognitive and perceptual mechanisms causing the voter error in question.

COGNITIVE MODEL

Based on Study 1 data, an initial cognitive model of the human voter was developed within the ACT-R framework. ACT-R is a unified theory of human cognition, a cognitive architecture that addresses both very low-level perceptual and motor capabilities of humans, as well as extremely high-level reasoning and decision making skills (ACT-R Research Group, 2009). The ACT-R framework has been used in a wide variety of research areas, ranging from basic visual search tasks (Tamborello, 2006; Tamborello & Byrne, 2006; Tamborello & Byrne, 2007) to modeling the approach and landing of pilots using synthetic vision systems (Byrne et al., 2004). To date however, ACT-R has not been used to model human voters. The current model therefore offers a meaningful opportunity to make a unique contribution to the field, by introducing modeling as a new technique in the assessment of voting system usability.

Method

Data Modeled

The model was constructed based on Study 1 data from those participants in the undirected information condition, as this condition was by far more ecologically valid and interesting than was the directed condition. Within the undirected condition, only data from those participants who saw a single race on screen one were utilized, as this screen layout was analogous to the ballot used in the real-world Sarasota election. Data for those participants who saw a single race on screen one were divided into two groups: participants who truly undervoted in the critical race (i.e. cast their ballot with no selection made for US Representative) versus those who did not skip the critical race at all. Restricting the data in this fashion was necessary because if people who initially skipped the critical race (but fixed their mistake prior to casting a vote) had been included, their screen times would have reflected the additional time spent on the screen when changing a response. The current model sought to compare search and selection strategies for voters their *first time* on the screens of interest, and did not address choice changes or error recovery at all. For the current model, only the first two screens of Study 1 were modeled (recall that screen two contained the critical US Representative race). It seemed probable that search and selection strategies might differ between participants based on the size of the difference in screen two times depending on whether the critical race was omitted or not. The size of this difference would not be explained solely by the additional encoding and motor movement required to select a candidate for that race.

Participants who truly undervoted in the critical race (who saw a single race on screen one and were in the undirected information condition) had a mean screen two time of 8.69 seconds (*SD* = 8.89). In contrast, the mean screen two time for those participants who did *not* skip the critical race (again, who saw a single race on screen one and were in the undirected condition) was much higher: 27.17 seconds (*SD* = 23.15). Mean screen one times for the two participant groups were 24.51 seconds (*SD* = 27.20) and 20.03 seconds (*SD* = 11.27), respectively. However, it must be noted that the relatively small number of true critical race undervotes was further reduced due to limiting data to only the undirected information condition and single race on screen one: data from only three participants remained after these specifications were met. For those participants who did not skip the critical race at all, data from 30 voters remained after the specifications were met. A ratio of 10 to 1 is clearly unbalanced; one would ideally have many more human data points against which the model could be compared. However, this problem actually reinforces the idea that cognitive modeling could be a particularly useful tool in the voting arena. In the future, a sufficiently complete model could easily examine effects that would normally require prohibitively large numbers of participants to do so in laboratory studies. That level of model complexity was well beyond the scope of the current project, which was intended solely to introduce cognitive modeling as a potential ballot design evaluation tool in the voting arena.

Design

The current model was designed to explore two different voting strategies that seemed plausible given the timing differences previously discussed: recalling a useful location from screen one and using it to direct visual search on screen two, versus reading screen two from top-to-bottom (or at least until a preferred candidate/party is encoded). These two strategies will be referred to as "recall" versus "read," and are discussed in more detail shortly. Two slightly different versions of the same basic model were initially implemented, with the short-term goal of determining whether use of a recall versus read strategy could account for the timing differences between those participants who undervoted in the critical race versus those who did not: the recall strategy resulted in a critical race undervote, whereas the read strategy did not. Obviously, a pressing future goal is the need to combine these two strategies into a single model: ideally, over the course of many model runs, the percentage of the time the model successfully executes the recall strategy (i.e. undervotes in the critical race) should approximate the Study 1 critical race undervote rate. The vast majority of the time, the read strategy would be selected over the recall strategy.

Currently however, slightly different versions of the same model are used to represent the two strategies. The two models start with identical declarative knowledge, and do not begin to diverge procedurally until they reach the second voting screen. Since participants in Study 1 (and Study 2) would have already used the navigation buttons to move forward from the instructions screen to the first voting screen, both models start with general knowledge about the navigation buttons (e.g. their purpose, color, and

screen location). Both models have basic knowledge about political parties (e.g. that "DEM" is a party abbreviation) and offices (e.g. that "senator" is an office name), and start with the same goal: completing a ballot. When either model "sees" the LISP window for screen one, a cycle of productions fires repeatedly that find, attend, and code the highest unexamined text on the screen, checking to see whether the newly examined text is an office name. If the text is not an office name, the models again cycle through those productions until an office name is found. Each time a new piece of text is examined, the models encode its screen location and "usefulness" (i.e. an office name is useful when trying to complete a ballot, whereas the general ballot header is not).

Once the first office is found, the models cycle through a similar series of productions that find, attend, and encode the top-most candidate name for that office, followed by analogous productions that find, attend, and encode the corresponding party abbreviation. The models cycle through these alternating productions until all candidate options have been encoded, at which time a retrieval request is made for the location of the "preferred" party. In each case, it was specified as being the Democratic party, since many participants voted a straight-party Democratic ticket. Most importantly, the three participants who truly undervoted in the critical race all selected the Democratic candidate for races one and three (having undervoted in race two). Once the models remembered their party preference, they located said party on the screen and clicked upon it with the mouse. The models verify that a checkmark indeed appears after they click the party, then resume searching the screen for another office. If another office is not found (which was the case on screen one, which presented only the Senator race), the models

make a retrieval request for the location of the forward navigation button, find it, and then click it. It is only at this point that the two models begin to diverge in terms of implementing a recall versus read strategy. The recall model makes a retrieval request for the location of a previously examined "useful" object, i.e. the location of the selection made on screen one, followed by the standard find, attend, encode, and click productions. In contrast, the read model implements the same strategy used on screen one: reading the screen from top to bottom before selecting a candidate. Both models were completely symbolic and deterministic, meaning that there was no variability in model candidate selections or model reaction times (i.e. mouse-click times).

Materials

The DRE ballot with which the model "voted" was similar to the DRE used by participants in terms of screen locations for offices, candidates, parties, and navigation buttons (Figures 31 and 32) and was programmed in LISP.

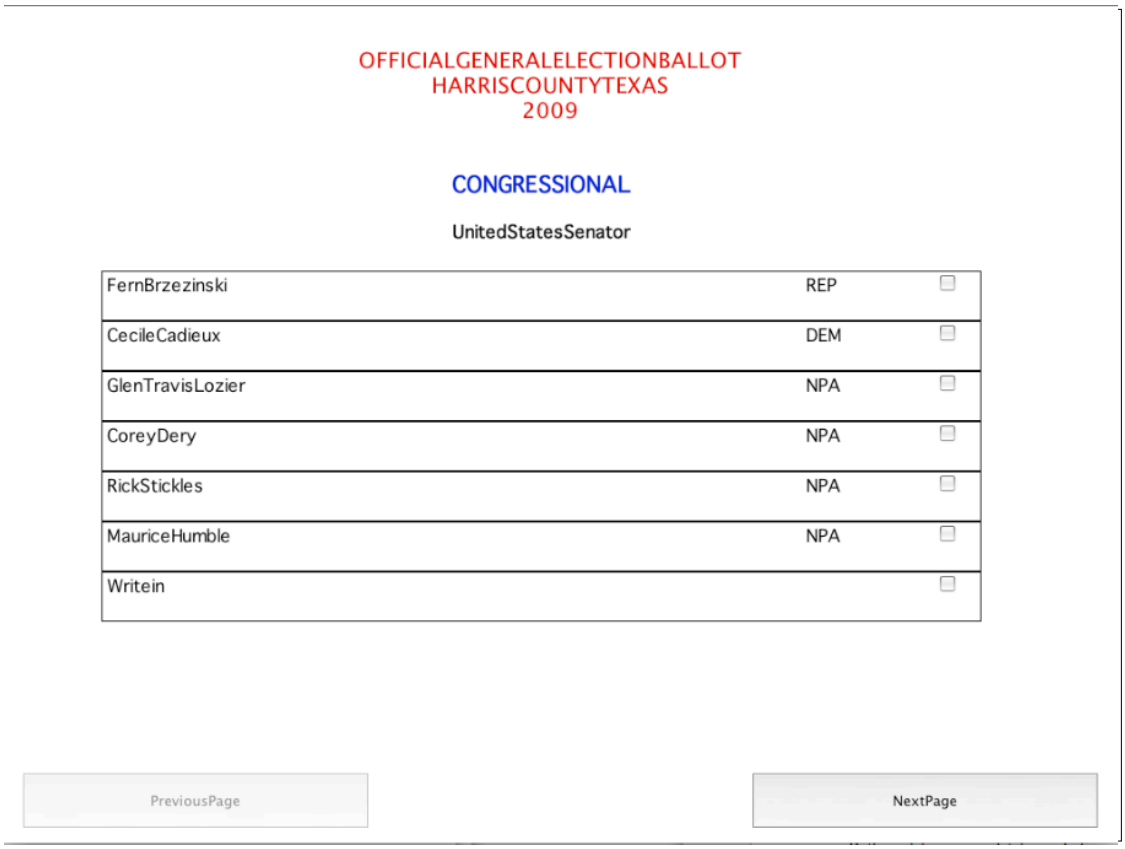


Figure 31. Model's mock DRE ballot, screen one: LISP version of first DRE screen (single race only) seen by participants in Study 1.

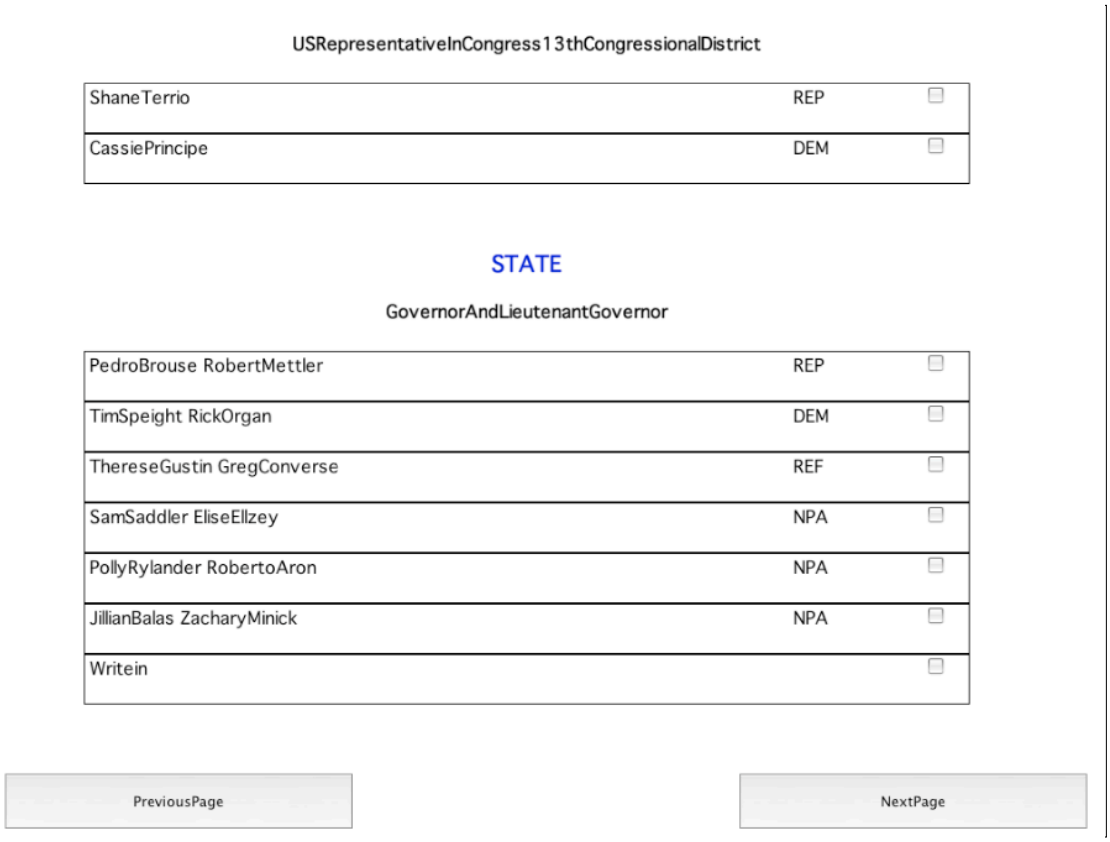


Figure 32. Model's mock DRE ballot, screen two: LISP version of second DRE screen (critical race screen) seen by participants in Study 1.

Results

Since the initial model was deterministic, each model run resulted in the same selections and times. Therefore, the standard measures of fit were not applicable; computing the correlation between model and human data, as well as the deviation between them, would only be meaningful with a non-deterministic model. As previously discussed, implementing this is a pressing future goal, as is modeling the condition in which participants saw two races on the first voting screen. Overall, the use of the recall

versus read strategies resulted in a qualitative match with the Study 1 human data in terms of the difference in times between screens one and two. However, the model was consistently faster than were the human participants. The qualitative predictions for Study 1 were encouraging. Yet a major idea behind implementing the two model versions was trying to assess the validity of differing strategies: while the recall strategy worked correctly for Study 1, it would not predict the results seen in Study 2. In Study 2, there was no effect of race location on undervotes in the critical Attorney General race. Had the model strategy been correct in using the location of a previously useful item to direct visual search on the following screen, participants who saw a lone race at the bottom of the screen should have been significantly less likely to miss the critical race on the subsequent screen. While the model in its current implementation (i.e. two separate models) is by no means intended to be a complete representation of a human voter, it certainly remains a useful starting point upon which a more sophisticated model should be built.

Future Directions

Once a non-deterministic model were successfully implemented, the fit between model data and human data from Study 1 could be more accurately assessed using metrics of correlation (R^2) and mean absolute deviation (MAD). Model parameters fit from Study 1 data would ideally be re-used for modeling Study 2 data in more depth. The most difficult test for any model is to make accurate predictions when facing novel stimuli. In the future, the model should be more flexible in handling novel stimuli, i.e. it

should be general enough to process novel stimuli without much rewriting of the model code. Assessing the accuracy of model predictions for the screens that differ between Study 1 and Study 2 should therefore be a future focus of model evaluation for Study 2. The fundamental model strategy of "recall" seems incorrect—or at least incomplete—given Study 2 results, so model predictions for both Study 1 and Study 2 must clearly be reevaluated after future iterations of the model.

Determining whether a model's predictions are accurate is not a simple yes or no question. There are increasingly rigorous levels of accuracy in model predictions. At the lowest level, the model should predict qualitative results, such as the ordering of group means seen in the human data. Predicting the ordering of condition means is usually the highest level of accuracy expected via heuristic evaluation in the absence of any user testing. Therefore, any increase in accuracy of model prediction beyond the ordering of group means would represent a unique benefit of cognitive modeling, providing information unlikely to be obtained by heuristic evaluation alone.

A level above predicting the ordering of group means would be approximate quantitative predictions. This level of accuracy would be considered obtained if the future model predictions (again, for DRE error rates and ballot completion times only) were to fall within 20% of the actual human data values. Group means should be evaluated for the various experimental conditions (for example, when two races were seen on screen one, rather than just modeling the single race condition in Study 1), as well as the degree to which the model captures trends in the data. As previously mentioned, R^2 and MAD would be used as evaluation metrics.

Finally, beyond approximate quantitative predictions, the highest possible level of accuracy would be perfect quantitative predictions of absolute values. Every aspect of the model's predictions would have to be exactly mirrored in the human data. Perfect quantitative prediction is not among the goals of even long-term future models, but achieving the ordering of group means would be expected, and ideally some level of approximate quantitative prediction as well. Irrespective of the degree of accuracy seen in future model predictions, areas where the model and human data deviate would continue to be informative. Disparities between model predictions and human data may help pinpoint limitations of the model itself, limitations in understanding human voters, or limitations in the data. Such identification of possible knowledge gaps shall always be necessary to help direct future research.

Implications of Model Predictions for Study 2

To the extent that model predictions of human performance for Study 2 were at least qualitatively correct, it would have indicated a degree of understanding of the cognitive processes of the voter, and how such cognitive mechanisms are affected by various aspects of interface design. Model predictions for Study 2 were not qualitatively correct, meaning that at least as measured by critical race undervote errors, people were not utilizing the "recall" strategy posited by the model. Although disappointing, this point of disparity between model predictions and Study 2 human data remains informative, helping to elucidate a specific area where our understanding of the interaction between voter and ballot (and perhaps between any user with a similar interface) may be lacking.

While it is possible that the small number of true critical race undervote errors seen in Study 2 were insufficient to ascertain any effects of top versus bottom race location, it is more likely that the model was not sophisticated enough to capture the subtleties of error generation seen in the human data. Future research should investigate disparities between model prediction and human performance, in an effort towards a more complete understanding of the human voter.

STUDY 2

Method

Participants

Except were otherwise specified, the experimental method for Study 2 was nearly identical to that used for Study 1. As the real intent of Study 2 was to collect human behavioral data against which model predictions could be compared, far fewer participants were used than in Study 1. For Study 2, a total of 50 participants were recruited from the larger Houston population: 25 female and 25 male. Ages ranged from 19 to 79 years, with a mean age of 46.7 (*SD*=19.6 years), and median age of 43 years. As in Study 1, participants were diverse in terms of education, ethnicity, income, and voting experience.

Design

A mixed design with one within-subjects variable and two between-subjects variables was used. The within-subjects variable was "DRE versus bubble ballot," since each participant completed the same ballot two times: once with a DRE and once with a paper ballot. Since the use of punch cards in Study 1 had no significant effect on the main variable of interest (critical race omissions), the non-DRE voting method for Study 2 was changed to paper ballots rather than punch cards. Not only are paper ballots easier to score than punch cards, but most importantly, paper ballots are more ecologically valid: they resemble provisional or absentee ballots, and in cases where DREs have been decertified, states have resumed use of optical scan paper ballots (rather than punch cards or lever machines).

The first between-subjects variable was information condition: 38 participants were in the undirected information condition, while only 12 participants were in the directed information condition. A larger number of participants were used in the undirected condition since it was the only information condition of interest for model creation, prediction, and evaluation. In the undirected information condition for Study 2, all participants received the same instructions: vote consistently and as you would in a real election. Recall that in Study 1, there was no effect of changing instructions from those just described to "special" instructions telling people to make a choice in *every* race, hence the decision not to include the "special" instructions in Study 2. Whereas the directed condition in Study 1 instructed participants to intentionally omit certain races based on real-world roll-off rates (Nichols and Strizek, 1995), the directed condition in

Study 2 did not make use of roll-off. In other words, the sheet participants were given listing the selections to make while voting did not include any intentional omissions for Study 2.

The second between-subjects variable was race location, which referred to whether the lone race on the second voting screen (i.e. the screen immediately preceding the critical race screen) was located at the *top* versus the *bottom* of the screen (Figures 33 and 34, respectively; see Appendices for screenshots of the Study 2 DRE ballot in its entirety).

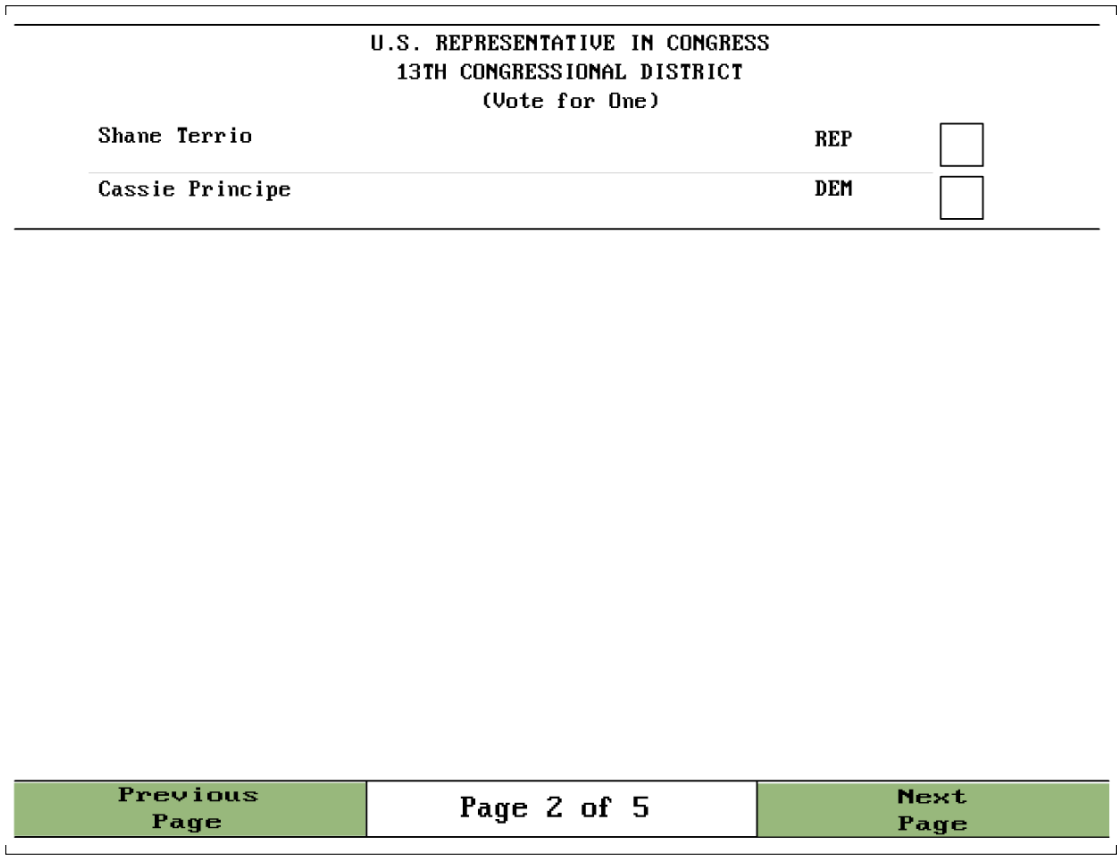


Figure 33. Screenshot of "top" race location condition for Study 2 DRE.

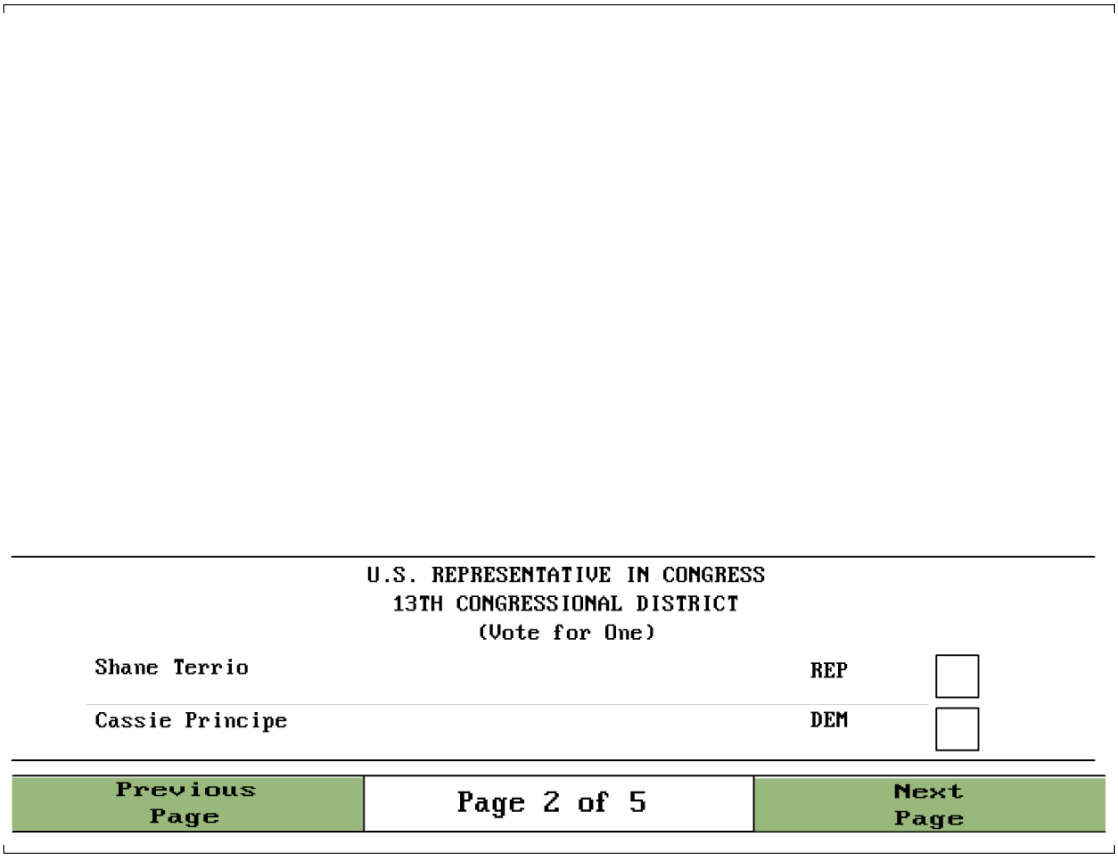


Figure 34. Screenshot of "bottom" race location condition for Study 2 DRE.

Since the manipulation of header highlighting in Study 1 did not predict critical race omissions, highlighting was removed entirely in Study 2. An additional change in Study 2 was the introduction of a "Charlotte County" screen (Figure 35), which was seen by all participants.

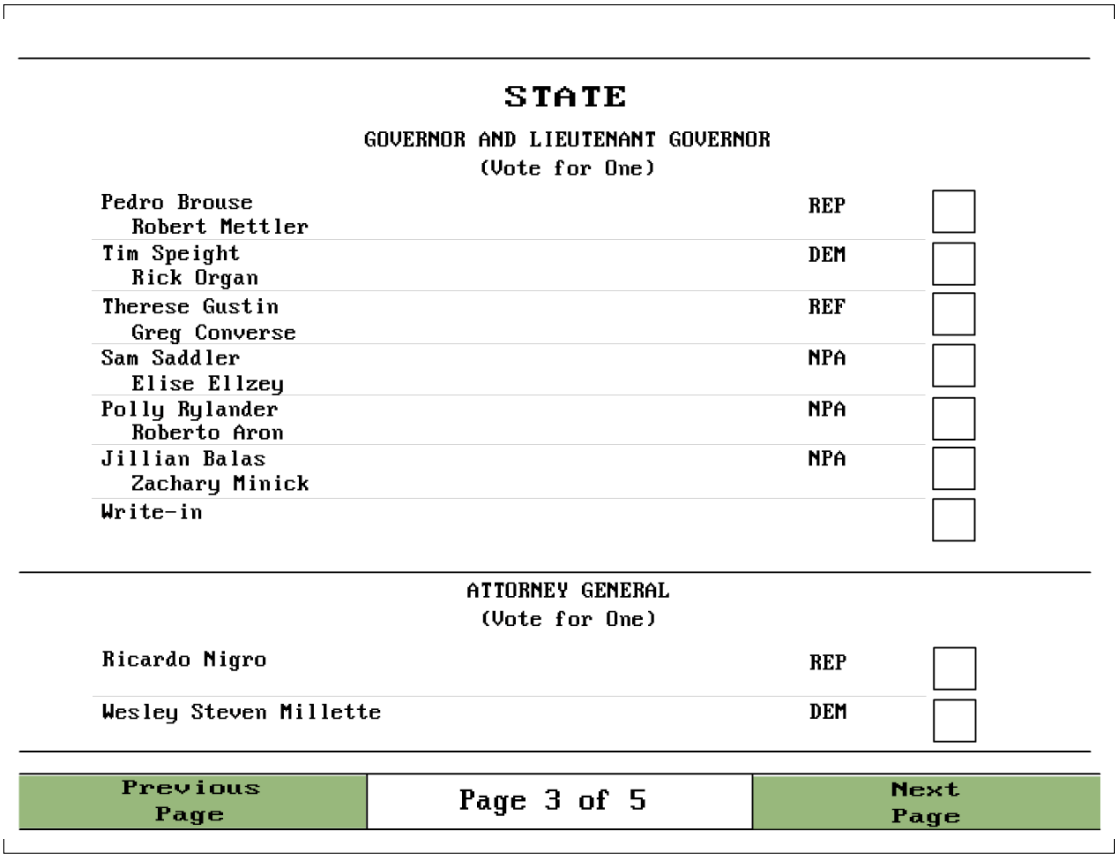


Figure 35. Screenshot of critical race screen for Study 2 DRE.

Recall that the critical race in Study 1 was the US Representative contest, seen on the second DRE voting screen. In Study 2, the critical race was the Attorney General contest, seen on the third voting screen. This screen mimicked the problematic two-race screen used in Charlotte County in 2006, where the Governor/Lieutenant Governor race and the Attorney General race were presented on the same screen. The distribution of screen space between the two races was reversed from that on the Sarasota ballot; whereas the top race on the Sarasota ballot took up considerably less screen space than did the bottom race, the opposite was true of the Charlotte County ballot (Figure 36).

Figure 36. Original critical race DRE screens from 2006 General Election in Sarasota County (top) and Charlotte County (bottom).

Materials and Procedure

The paper ballot stands and DRE hardware for Study 2 were those used in previous research. With the exception of the Study 2 DRE ballot design changes previously described, DREs were similar in appearance to those used in Study 1 (see Appendices for individual screen shots from Studies 2 and 1, respectively).

Overall, procedures for Study 2 were nearly identical to those used in Study 1. Study 2 procedures differed in that paper ballots rather than punch cards were used, and all participants in the undirected information condition received the same instructions (recall that manipulation of instructions in the undirected information condition had no effect in Study 1).

Participants again began by giving informed consent and receiving instructions. Participants in the undirected information condition were then offered a voter guide, while participants in the directed condition were given a sheet of paper listing all the selections they were to make while voting. Participants again voted first on a DRE, then completed a SUS questionnaire for that method. Participants then completed a paper ballot and corresponding SUS questionnaire. Participants in the undirected information condition then completed an exit interview. All participants completed a final survey packet, and were then debriefed and paid.

Results

Effectiveness: Critical Race Results

For Study 2, only those results directly relevant to the primary variable of interest (critical race omissions) shall be discussed. Data from all 50 participants were included in the following critical race analyses. On the DRE, eight of 50 voters skipped the Attorney General race their first time on that screen, resulting in an overall initial omission rate of **16.00%**. Of those eight participants, six corrected their critical race undervote error before casting their final ballot; five of those six voters made this correction pre-review screen, while only one voter corrected the mistake post-review screen. Interestingly, two of those six voters were in the *directed* information condition: even when they had a piece of paper listing the Attorney General race in front of them, they skipped it nonetheless. Two of the eight voters who initially skipped the Attorney General race never corrected their mistake, resulting in an overall final undervote rate of **4.00%**. As the model was designed to be representative of voters in the undirected information condition only, critical race results were also computed for the undirected condition alone; the percentages were similar regardless of whether all participants (N=50) or just those in the undirected information condition (n=38) were used (Table 4). Therefore, data from all participants were included when predicting critical race omissions in the following logistic regression.

Table 4. Study 2 critical race results.

	Initial Omissions		Final Undervotes	
	Count	Percentage	Count	Percentage
All participants (N = 50)	8	16.00%	2	4.00%
Undirected condition only (N = 38)	6	15.79%	2	5.26%

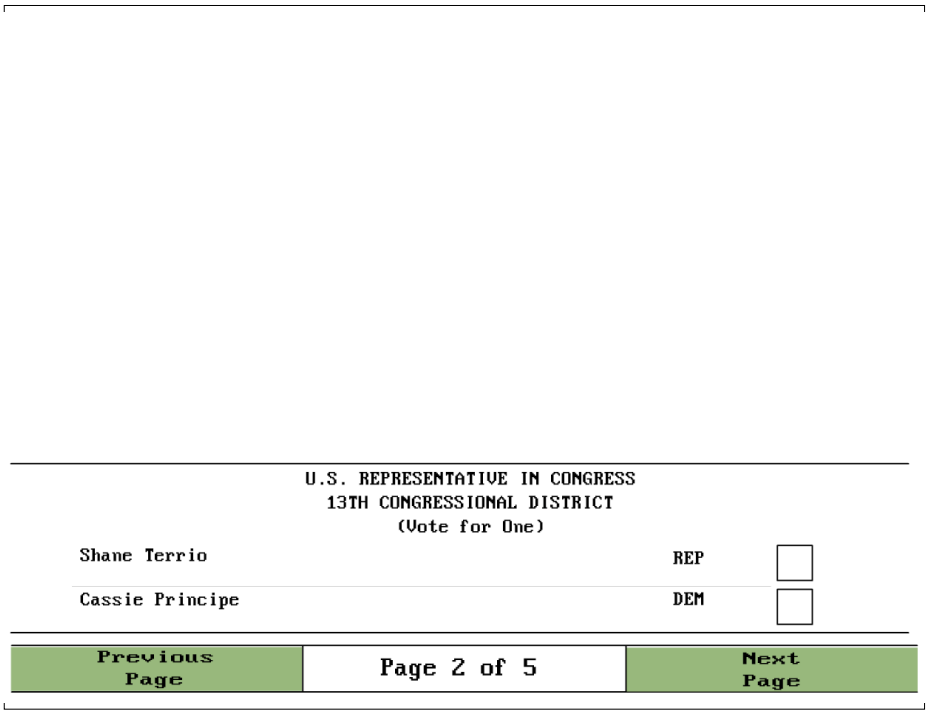
Critical race initial omissions were regressed on race location, education, and age (age as a continuous variable). All two-way interaction terms were tested initially in the first block of a logistic regression: none were significant. All main effects were then similarly evaluated: the model was not significant overall (Chi-square(5, $N = 50$) = 1.61, $p = .90$), accounting for only 5.4% of the variance in predicting whether participants skipped the critical race. The main effect of education was not reliable ($p = .81$), nor were the main effects of age ($b = .01$, $p = .78$), information condition ($b = -.14$, $p = .88$), or race location ($b = .72$, $p = .38$).

Discussion

Despite the fact that the distribution of screen real estate was reversed between the critical race screens in Study 1 and Study 2 (refer back to Figure 36 for original screen shots), participants in Study 2 nonetheless skipped the critical Attorney General race in disturbingly high numbers. Although the Study 2 critical race final undervote rate was much lower than the residual vote rate for that race during the actual election (4.00% versus 20.9%, respectively), the critical race initial omission rate was more comparable: 16.00% in Study 2 versus 20.9% in the actual election. This difference of 4.9% bears further discussion, since one would have expected the Study 2 critical race initial omission rate to be *greater* than the real-world residual vote rate given the pattern of results seen in Study 1. (Recall that the critical race initial omission rate of 19.71% in Study 1 was *greater* than the 13.9% residual vote rate for that race during the actual election in Sarasota County.) There may be additional explanations for this Study 2 result, but the following three points seem most relevant. (1) The number of participants run in Study 2 was far fewer than the number run in Study 1. Obviously, when looking for a specific type of error that is not consistently produced by participants, the probability of finding such an error may vary across different samples. (2) Header highlighting was removed entirely in Study 2, whereas it was present for half of all participants in Study 1. Although header highlighting did not significantly predict critical race omissions in Study 1, it did interact with other variables to impact error rates. It is quite probable that Study 2 results may have differed had original Charlotte County highlighting been used. (3) One would normally expect the residual vote rate for the Attorney General race to be greater

than the residual vote rate for the US Representative race, simply because the former is further "down-ballot" than is the latter. Recall that the critical race initial omission rate in Study 2 (as in Study 1) was a *true* measure of voter error, and specifically included *only* undervote errors. In contrast, the real-world residual vote rate is comprised not only of undervote errors, but also of intentional omissions and overvotes.

Overall, Study 2 was successful in replicating a real-world phenomenon from the 2006 General Election in Charlotte County, Florida: the abnormally large number of people who missed the critical Attorney General race relative to other nearby races. However, the manipulation of race location on the screen preceding the critical race did not effect critical race undervotes. This manipulation was designed primarily to test model predictions, specifically those made by the model strategy of recalling a useful location from a preceding screen to direct visual search strategy on a subsequent screen. If this model strategy were indeed correct (i.e. correct in the sense that it accurately represented human strategy), one would have expected fewer critical race undervote errors when the lone race on the preceding screen was at the bottom of the screen rather than the top, as this bottom race location would be deemed "useful" and used to inform visual search on the following critical race screen (Figure 37).



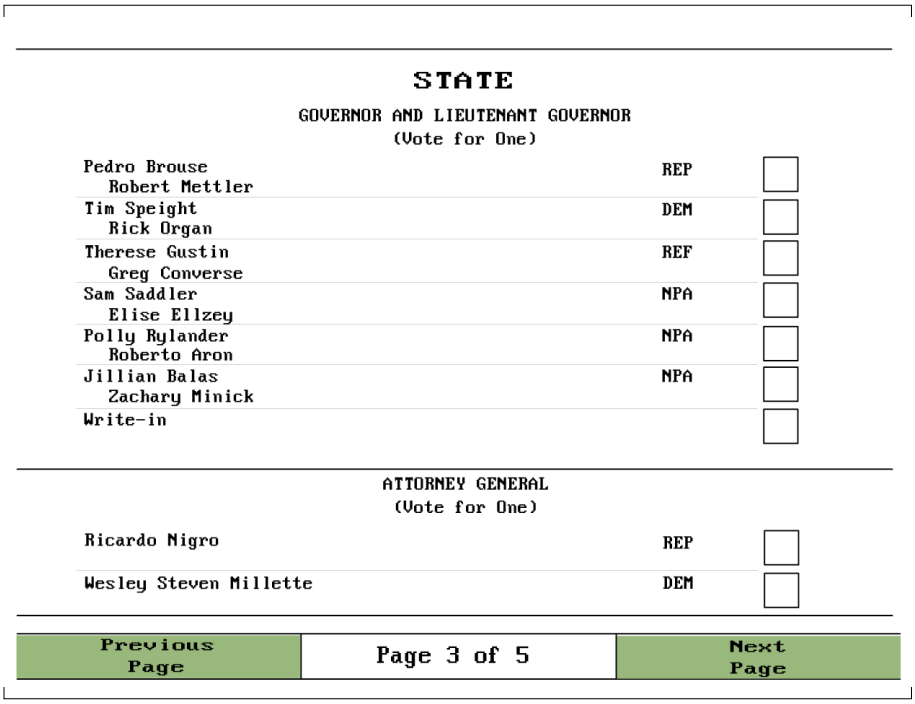


Figure 37. "Bottom" race location (top figure) and critical race screen (bottom figure) for Study 2 DRE.

GENERAL CONCLUSIONS

Theoretical and applied contributions of current studies

Although usability and human factors are increasingly being recognized as an important facet of voting systems, laboratory and field data on this topic are both relatively limited as of yet. The current studies contribute significantly to the voting research community in addressing this dearth of literature, by providing data from experiments run in well-controlled laboratory settings. In addition to yielding standard behavioral data, these studies offer a very unique contribution to the voting community: introduction of cognitive modeling as a potential research method in this arena. Not only is this informative for the voting community, but should also be of interest for the field of cognitive modeling and the broader HCI/HF community.

There are intriguing practical long-term uses for a sufficiently complete cognitive model of voters. In the future, a cognitive model of human voters could be used both as a pre- and post-election tool to evaluate effects of interface design decisions on election results. Pre-election, it could be used to help evaluate proposed ballot layouts. Having the model vote on a mocked-up version of the ballot could help predict—and ideally prevent—future performance problems, such as large undervote rates and systematic selection of a candidate other than the one intended. Rather than leaving it up to election officials to assess a ballot design, who are often overworked and do not necessarily have HCI/HF training, having an empirical means of evaluating ballot designs would be preferred. Cognitive modeling could also be used as a post-election forensics tool, to help evaluate claims that poor ballot design unfairly influenced specific election outcomes,

perhaps even reducing the length and cost of lawsuits by candidates contesting election results on a basis of ballot design.

In addition to the voting-specific applications just discussed, these studies could contribute to the broader field of HCI/HF overall. As technology continues to advance and computing becomes even more ubiquitous, people are performing an increasing number of tasks online. Banking, shopping, photo sharing, and social networking are just a few examples. This increasing reliance on the internet is not unique to individuals, but can be seen in a wide variety of organizations as well. At academic institutions, students apply for classes and check grades online, instructors distribute course materials electronically, etc. Businesses and branches of government are increasingly offering their services and products online. In all of these cases, users must interact with an electronic interface to accomplish their goals. Although the exact nature of the task for which it is intended must clearly influence design decisions, there remain issues common to all interfaces. Which design characteristics can help ensure users find the information necessary to complete a task? When is the use of highlighting beneficial versus detrimental? How much experience with an interface is needed before users expect it to work a certain way? Which interface design characteristics most influence human performance, and what are the cognitive mechanisms underlying such performance effects? Such questions are not unique to voting systems, and have important theoretical and practical implications for the larger HCI/HF community as well. Knowing which design features are most problematic is the first step to designing a better interface. In many modern-day scenarios, ranging from military cockpits to home computer desktops,

it simply is not feasible to present only a single choice/task at a time. The vast majority of the time, screen real estate is limited and the need for information is high; knowing the relative effects of different design features, such as highlighting and screen layout, would clearly be beneficial for a myriad of applications. While studies of highlighting and highlighting validity have certainly been done before, such research has not typically been applied to the task of voting. The current dissertation provides much-needed behavioral data upon which to base future theory and research in voting system usability, while introducing cognitive modeling as a new interface evaluation tool in the voting arena. Perhaps most importantly, these studies directly address an important real-world empirical question: what happened in the 2006 Sarasota election?

REFERENCES

Amunson, J. R., & Hirsch, S. (2008). The case of the disappearing votes: Lessons from the Jennings v. Buchanan Congressional Election Contest. *William & Mary Bill of Rights Journal, 17,* 397-422.

Brennan Center for Justice, Norden, L., Kimball, D. C., Quesenbery, W., & Chen, M. (2008). *Better Ballots.* Retrieved from http://brennan.3cdn.net/d6bd3c56be0d0cc861_hlm6i92vl.pdf.

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.

Byrne, M. D., Kirlik, A., Fleetwood, M. D., Huss, D. G., Kosorukoff, A., Lin, R., & Fick, C. S. (2004). A closed-loop, ACT-R approach to modeling approach and landing with and without synthetic vision system (SVS) technology. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 2111-2115). Santa Monica, CA: Human Factors and Ergonomics Society.

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

Doig, M. (2006, November 15). Analysis suggests undervote caused by ballot design. *Herald Tribune.* Retrieved from http://www.heraldtribune.com/apps/pbcs.dll/article?AID=/20061115/NEWS/611150751.

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

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

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

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

Greene, K. K. (2008). 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. Master's Thesis, Rice University, Houston, TX.

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., Niemi, R. G., Hanmer, M. J., Bederson, B. B., Conrad, F. C., & Traugott, M. W. (2008). Voting Technology. Washington, DC: Brookings Institution Press.

Jacobs, E. (2005). Spoiled ballots: Under and overvotes in the 2004 General Election in Montgomery County, Ohio 4. Retrieved from http://ohioelection2004.com/WordHost/montgomeryco4pctsellisjacobs.

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

Laskowski, S. J., Autry, M., Cugini, J., Killam, W., & Yen, J. (2004). Improving the usability and accessibility of voting systems and products. *NIST Special Publication 500-256.*

Mascher, A. L., Cotton, P. T., & Jones, D. W. (2010). Towards publishable event logs that reveal touchscreen faults. *Proceedings of the 2010 USENIX/ACCURATE Electronic Voting Technology Workshop/Workshop on Trustworthy Elections.* Washington, DC.

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

Nielsen, J. (2007). *Banner blindness in ballot design*. Sidebar to *Banner blindness: Old and new findings*. Retrieved from http://www.useit.com/alertbox/banner-blindness-ballot-design.html

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

Tamborello, F. P., II. (2006). Visual displays: The continuing investigations of the highlighting paradox. Master's Thesis, Rice University, Houston, TX.

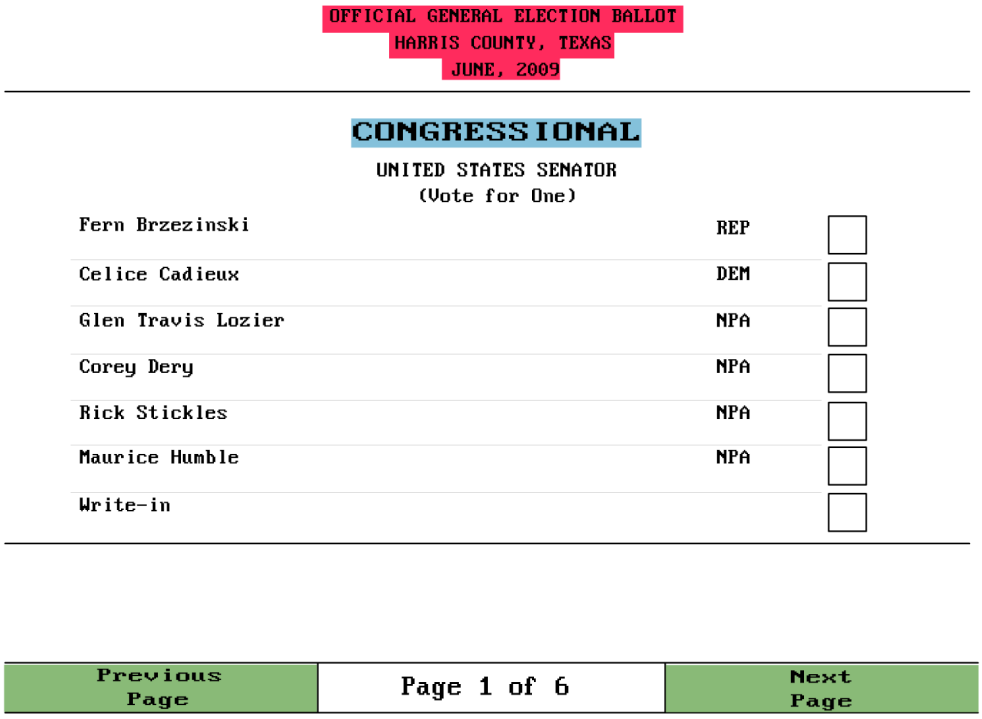
Tamborello, F. P., II, & Byrne, M. D. (2006). Adaptive but non-optimal visual search behavior in highlighted displays. *Proceedings of the Seventh International Conference on Cognitive Modeling.*

Tamborello, F. P., II, & Byrne, M. D. (2007). Adaptive but non-optimal visual search behavior in highlighted displays. *Cognitive Systems Research, 8 (3),* 182-191.

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

Wand, J. N., Shotts, K. W., Sekhon, J. S., Mebane, W. R., Herron, M. C., & Brady, H. E. (2001). The Butterfly Did it: The Aberrant Vote for Buchanan in Palm Beach County, Florida. *American Political Science Review, 95, 4,* 793-84.

APPENDIX A: STUDY 1 SCREENSHOTS



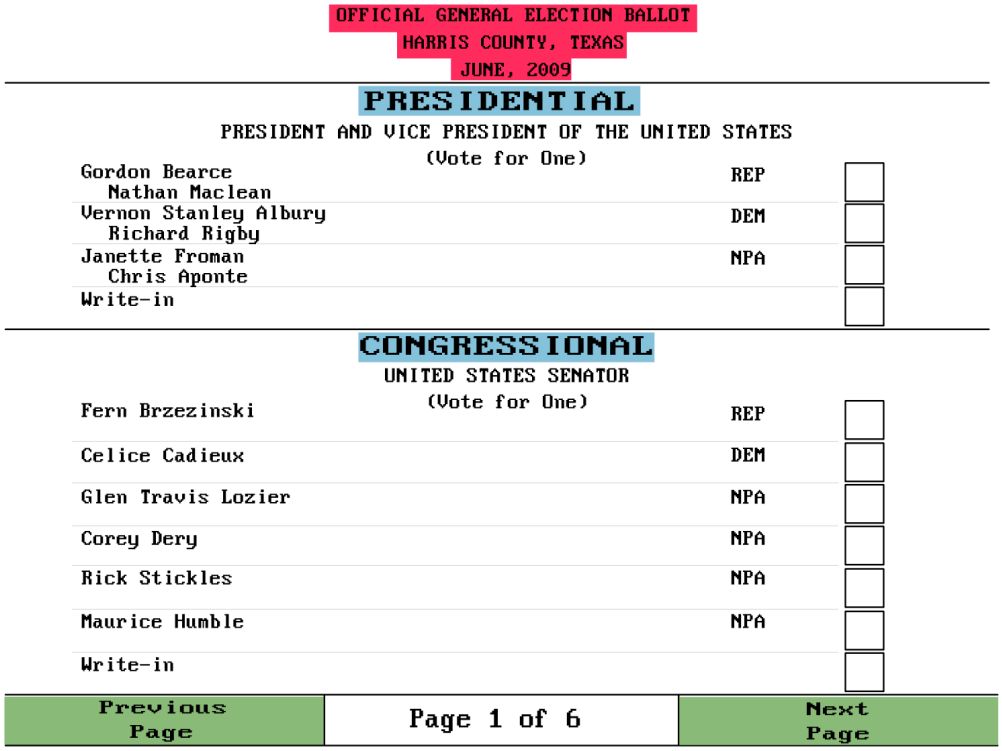
OFFICIAL GENERAL ELECTION BALLOT
HARRIS COUNTY, TEXAS
JUNE, 2009

CONGRESSIONAL

UNITED STATES SENATOR
(Vote for One)

Fern Brzezinski	REP	☐
Celice Cadieux	DEM	☐
Glen Travis Lozier	NPA	☐
Corey Dery	NPA	☐
Rick Stickles	NPA	☐
Maurice Humble	NPA	☐
Write-in		☐

Previous Page | Page 1 of 6 | Next Page



OFFICIAL GENERAL ELECTION BALLOT
HARRIS COUNTY, TEXAS
JUNE, 2009

PRESIDENTIAL

PRESIDENT AND VICE PRESIDENT OF THE UNITED STATES
(Vote for One)

Gordon Bearce Nathan Maclean	REP	☐
Vernon Stanley Albury Richard Rigby	DEM	☐
Janette Froman Chris Aponte	NPA	☐
Write-in		☐

CONGRESSIONAL

UNITED STATES SENATOR
(Vote for One)

Fern Brzezinski	REP	☐
Celice Cadieux	DEM	☐
Glen Travis Lozier	NPA	☐
Corey Dery	NPA	☐
Rick Stickles	NPA	☐
Maurice Humble	NPA	☐
Write-in		☐

Previous Page | Page 1 of 6 | Next Page

U.S. REPRESENTATIVE IN CONGRESS
13TH CONGRESSIONAL DISTRICT
(Vote for One)

Shane Terrio	REP	☐
Cassie Principe	DEM	☐

STATE

GOVERNOR AND LIEUTENANT GOVERNOR
(Vote for One)

Pedro Brouse Robert Mettler	REP	☐
Tim Speight Rick Organ	DEM	☐
Therese Gustin Greg Converse	REF	☐
Sam Saddler Elise Ellzey	NPA	☐
Polly Rylander Roberto Aron	NPA	☐
Jillian Balas Zachary Minick	NPA	☐
Write-in		☐

Previous Page | Page 2 of 6 | Next Page

ATTORNEY GENERAL
(Vote for One)

Ricardo Nigro	REP	☐
Wesley Steven Millette	DEM	☐

CHIEF FINANCIAL OFFICER
(Vote for One)

Petra Bencomo	REP	☐
Susanne Rael	DEM	☐

COMMISSIONER OF AGRICULTURE
(Vote for One)

Peter Varga	REP	☐
Mark Baber	DEM	☐

Previous Page | Page 3 of 6 | Next Page

LEGISLATIVE

STATE REPRESENTATIVE
71ST HOUSE DISTRICT
(Vote for One)

Tim Grasty	REP	☐
Write-in		☐

COUNTY

CHARTER REVIEW BOARD DISTRICT 1
(Vote for One)

Dan Plouffe	REP	☐
Derrick Melgar	DEM	☐

CHARTER REVIEW BOARD DISTRICT 2
(Vote for One)

Corey Behnke	REP	☐
Jennifer A. Lundeed	NPA	☐

CHARTER REVIEW BOARD DISTRICT 3
(Vote for One)

Dean Caffee	REP	☐
Gordon Kallas	DEM	☐

CHARTER REVIEW BOARD DISTRICT 4
(Vote for One)

Stanley Saari	REP	☐
Jason Valle	DEM	☐

CHARTER REVIEW BOARD DISTRICT 5
(Vote for One)

Howard Grady	REP	☐
Randy H. Clemons	DEM	☐

HOSITAL BOARD SOUTHERN DISTRICT SEAT 1
(Vote for One)

Deborah Kamps	REP	☐
Clyde Gayton Jr.	DEM	☐

Previous Page	Page 6 of 6	Next Page

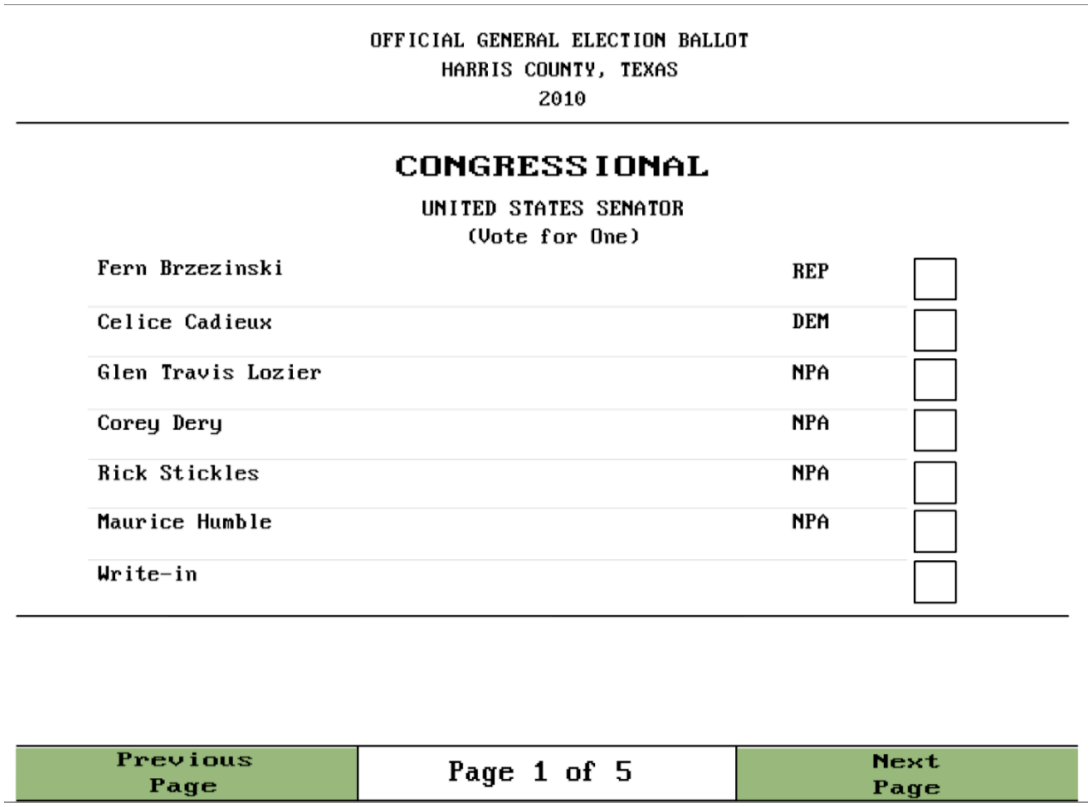
Instructions

Return to any contest
by touching the contest title.
To cast your ballot, click the
VOTE button on the next page.

PRESIDENT AND VICE PRESIDENT..... ☐ No Selection Made	STATE REPRESENTATIVE..... ☐ No Selection Made
UNITED STATES SENATOR..... ☐ No Selection Made	CHARTER REVIEW BOARD DISTRIC..... ☐ No Selection Made
U.S.REPRESENTATIVE IN CONGR..... ☐ No Selection Made	CHARTER REVIEW BOARD DISTRIC..... ☐ No Selection Made
GOVERNOR AND LIEUTENANT GOVE..... ☐ No Selection Made	CHARTER REVIEW BOARD DISTRIC..... ☐ No Selection Made
ATTORNEY GENERAL..... ☐ No Selection Made	CHARTER REVIEW BOARD DISTRIC..... ☐ No Selection Made
CHIEF FINANCIAL OFFICER..... ☐ No Selection Made	CHARTER REVIEW BOARD DISTRIC..... ☐ No Selection Made
COMMISSIONER OF AGRICULTURE..... ☐ No Selection Made	HOSPITAL BOARD SOUTHERN DIST..... ☐ No Selection Made

Previous Page	Summary Ballot	Next Page

APPENDIX B: STUDY 2 SCREENSHOTS



OFFICIAL GENERAL ELECTION BALLOT
HARRIS COUNTY, TEXAS
2010

CONGRESSIONAL

UNITED STATES SENATOR
(Vote for One)

Fern Brzezinski	REP	☐
Celice Cadieux	DEM	☐
Glen Travis Lozier	NPA	☐
Corey Dery	NPA	☐
Rick Stickles	NPA	☐
Maurice Humble	NPA	☐
Write-in		☐

Previous Page | Page 1 of 5 | Next Page

U.S. REPRESENTATIVE IN CONGRESS
13TH CONGRESSIONAL DISTRICT
(Vote for One)

Shane Terrio	REP	☐
Cassie Principe	DEM	☐

Previous Page | Page 2 of 5 | Next Page

U.S. REPRESENTATIVE IN CONGRESS
13TH CONGRESSIONAL DISTRICT
(Vote for One)

Shane Terrio	REP	☐
Cassie Principe	DEM	☐

Previous Page | Page 2 of 5 | Next Page

STATE

GOVERNOR AND LIEUTENANT GOVERNOR
(Vote for One)

Pedro Brouse Robert Mettler	REP	☐
Tim Speight Rick Organ	DEM	☐
Therese Gustin Greg Converse	REF	☐
Sam Saddler Elise Ellzey	NPA	☐
Polly Rylander Roberto Aron	NPA	☐
Jillian Balas Zachary Minick	NPA	☐
Write-in		☐

ATTORNEY GENERAL
(Vote for One)

Ricardo Nigro	REP	☐
Wesley Steven Millette	DEM	☐