Now Do Voters Notice Review Screen Anomalies?

A Look at Voting System Usability

Bryan A. Campbell
Michael D. Byrne
Department of Psychology
Rice University
Houston, TX
bryan.campbell@rice.edu
byrne@acm.org
http://chil.rice.edu/
Overview

• Background
  • Usability and security
  • Previous research on review screen anomaly detection

• Methods
  • New experiment on anomaly detection

• Results
  • Improved detection
  • Replication of some previous findings
  • New findings

• Discussion
Consider the amount of time and energy spent on voting system security, for example:

- California’s Top-to-Bottom review
- Ohio’s EVEREST review
- Many other papers past and present EVT/WOTE

This despite a lack of conclusive evidence that any major U.S. election has been stolen due to security flaws in DREs

- Though of course this *could* have happened

But we know major U.S. elections have turned on voting system usability
OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

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

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

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

(GREEN)
RALPH NADER - PRESIDENT
WINONA LA DUKE - VICE PRESIDENT

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

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

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

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

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

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

WRITE-IN CANDIDATE
To vote for a write-in candidate, follow the directions on the long stub of your ballot card.
U.S. REPRESENTATIVE IN CONGRESS  
13TH CONGRESSIONAL DISTRICT  
(Vote for One)

Vern Buchanan  
Christine Jennings

| STATE |
| GOVERNOR AND LIEUTENANT GOVERNOR |
| (Vote for One) |

| Charlie Crist  |
| Jeff Kottkamp  |
| Jim Davis  |
| Daryl L. Jones  |
| Max Linn  |
| Tom Macklin  |
| Richard Paul Dembinsky  |
| Dr. Joe Smith  |
| John Wayne Smith  |
| James J. Kearney  |
| Karl C.C. Behm  |
| Carol Castagnero  |
| Write-In  |
There are numerous other examples of this

• See the 2008 Brennan Center report

This is not to suggest that usability is more important than security

• Though we’d argue that it does deserve equal time, which has not been the case

Furthermore, usability and security are intertwined

• The voter is the first line of defense against malfunctioning and/or malicious systems

• Voters may be able to detect when things are not as they should be

✦ The oft-given “check the review screen” advice
Other usability findings from our previous work regarding DREs vs. older technologies

- Voters are not more accurate voting with a DRE
- Voters are not faster voting with a DRE
- However, DREs are vastly preferred to older voting technologies

But do voters actually check the review screen?

- Or rather, how closely do they check?
- Assumption has certainly been that voters do

Everett (2007) research

- Two experiments on review screen anomaly detection using the VoteBox DRE
Below are the choices you have made. If you would like to make changes, click on the race you would like to change.

If you do not want to make changes, click the 'Next Page' button to go to Step 4.

**Your vote will not be recorded unless you finish step 4.**

<table>
<thead>
<tr>
<th>Position</th>
<th>Candidate</th>
<th>Position</th>
<th>Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>Vernon Stanley Albur</td>
<td>Judge on Court of Criminal Appeals</td>
<td>Dan Plouffe</td>
</tr>
<tr>
<td>Vice President</td>
<td>Richard Rigby</td>
<td></td>
<td></td>
</tr>
<tr>
<td>United States Senator</td>
<td>None</td>
<td>District Attorney of Harris County</td>
<td>None</td>
</tr>
<tr>
<td>US House of Representative</td>
<td>None</td>
<td>County Treasurer of Harris County</td>
<td>None</td>
</tr>
<tr>
<td>Governor of Texas</td>
<td>None</td>
<td>Sheriff of Harris County</td>
<td>None</td>
</tr>
<tr>
<td>Lieutenant Governor of Texas</td>
<td>Cassie Principe</td>
<td>County Tax Assessor of Harris Country</td>
<td>None</td>
</tr>
<tr>
<td>Attorney General of Texas</td>
<td>Tim Speight</td>
<td>Justice of the Peace of Harris County</td>
<td>None</td>
</tr>
<tr>
<td>Comptroller of Public Accounts</td>
<td>Greg Converse</td>
<td>County Judge of Harris County</td>
<td>None</td>
</tr>
<tr>
<td>Commissioner of General Land Office</td>
<td>Sam Saddler</td>
<td>Proposition 1:</td>
<td>Yes</td>
</tr>
<tr>
<td>Commissioner of Agriculture</td>
<td>Roberto Aron</td>
<td>Proposition 2:</td>
<td>No</td>
</tr>
<tr>
<td>Railroad Commissioner of Texas</td>
<td>Jillian Balas</td>
<td>Proposition 3:</td>
<td>Yes</td>
</tr>
<tr>
<td>State Senator of Texas</td>
<td>None</td>
<td>Proposition 4:</td>
<td>No</td>
</tr>
<tr>
<td>State Representative of Texas</td>
<td>None</td>
<td>Proposition 5:</td>
<td>None</td>
</tr>
<tr>
<td>State Board of Education</td>
<td>Mark Baber</td>
<td>Proposition 6:</td>
<td>None</td>
</tr>
<tr>
<td>Presiding Judge on Texas Supreme Court</td>
<td>Tim Grasty</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Click to go back to previous race

Click to go to Step 4: Record your vote

Next Page→
Everett (2007)

First study

- Two or eight entire contests were added or subtracted from the review screen

Second study

- One, two, or eight changes were made to the review screen
- Changes were to an opposing candidate or an undervote and appeared on the top or bottom of the ballot

Results

- First study: 32% noticed the anomalies
- Second study: 37% noticed the anomalies
Also examined what other variables did and did not influence detection performance

Affected detection performance:

- Time spent on review screen
  - Causal direction not clear here
- Whether or not voters were given a list of candidates to vote for
  - Those with a list noticed more often

Did not affect detection performance:

- Number of anomalies
- Location on the ballot of anomalies
Everett (2007) Limitations

- Participants were never explicitly told to check the review screen.
  - Would simple instructions increase noticing rates?
- The interface did little to aid voters in performing accuracy checks
  - Was there too little information on the screen?
Explicit instructions

- Voting instructions, both prior to and on the review screen, explicitly warned voters to check the accuracy of the review screen.

Review screen interface alterations

- Undervotes were highlighted in a bright red-orange color.
- Party affiliation markers were added to candidate names on the review screen.
Below are the choices you have made. If you would like to make changes, click on the race you would like to change. **Please be sure to review your choices and correct any mistakes *before* casting your ballot.**

If you do not want to make changes, click the ‘Next Page’ button to go to Step 4.

**Your vote will not be recorded unless you finish step 4.**

<table>
<thead>
<tr>
<th>Position</th>
<th>Candidate(s)</th>
<th>Party</th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>Gordon Bearer</td>
<td>R</td>
</tr>
<tr>
<td>Vice President</td>
<td>Nathan MacLean</td>
<td>R</td>
</tr>
<tr>
<td>United States Senator</td>
<td>Fern Brzezinski</td>
<td>D</td>
</tr>
<tr>
<td>US House of Representative</td>
<td>Pedro Brouse</td>
<td>R</td>
</tr>
<tr>
<td>Governor of Texas</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Lieutenant Governor of Texas</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Attorney General of Texas</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Comptroller of Public Accounts</td>
<td>Therese Gustin</td>
<td>I</td>
</tr>
<tr>
<td>Commissioner of General Land Office</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Commissioner of Agriculture</td>
<td>Roberto Aron</td>
<td>D</td>
</tr>
<tr>
<td>Railroad Commissioner of Texas</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>State Senator of Texas</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>State Representative of Texas</td>
<td>Petra Bencomo</td>
<td>R</td>
</tr>
<tr>
<td>State Board of Education</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Presiding Judge on Texas Supreme Court</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Judge on Court of Criminal Appeals</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>District Attorney of Harris County</td>
<td>Jennifer A. Lundiee</td>
<td>D</td>
</tr>
<tr>
<td>County Treasurer of Harris County</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Sheriff of Harris County</td>
<td>Stanley Saari</td>
<td>R</td>
</tr>
<tr>
<td>County Tax Assessor of Harris County</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Justice of the Peace of Harris County</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>County Judge of Harris County</td>
<td>Dan Atchley</td>
<td>R</td>
</tr>
</tbody>
</table>

**Important**

Click to go back to previous contest

Click to go to Step 4: Record your vote
108 voters participated in our mock election

- Recruited from the greater Houston area via newspaper ads, paid $25 for participation
- Native English speakers 18 years of age or older
- Mean age = 43.1 years (SD = 17.9); 60 female, 48 male
- Previous voting experience: mean number of national elections was 5.8, mean non-national elections was 6.3
- Self-rated computer expertise mean of 6.2 on a 10-point Likert scale
Design: Independent Variables

- Number of anomalies
  - Either 1, 2, or 8 anomalies were present on the review screen
- Anomaly type
  - Contests were changed to an opposing candidate or to an undervote
- Anomaly location
  - Anomalies were present on either the top or bottom half of the ballot
Design: Independent Variables

Information condition

- Undirected: Voter guide, voters told to vote as they wished
- Directed: Given list of candidates to vote for, cast a vote in every race
- Directed with roll-off: Given a list of candidates to vote for, but instructed to abstain in some races

Voting system

- Voters voted on the DRE and one other non-DRE system

Other system

- Voters voted on either a bubble-style paper, lever machine, or punch card voting system
Design: Dependent Variables

- **Anomaly detection**
  - Voters, by self-report, either noticed the anomalies or they did not
  - Also, self-report on how carefully the review screen was checked

- **Efficiency**
  - Time taken to complete a ballot

- **Effectiveness**
  - Error rate

- **Satisfaction**
  - Subjective SUS scores
Design: Error Types

- Wrong choice errors
  - Voter selected a different candidate
- Undervote errors
  - Voter failed to make a selection
- Extra vote errors
  - Voter made a selection when s/he should have abstained
- Overvote errors
  - Made multiple selections (DRE and lever prevent this error)

Also, voters in the undirected condition could intentionally undervote, though this is not an error
  - Raises issue of true error rate vs. residual error rate
Results: Anomaly Detection

- **50%** of voters detected the review screen anomalies
  - 95% confidence interval: 40.1% to 59.9%
  - Clear improvement beyond Everett (2007), but still less than ideal

So, what drove anomaly detection?

- Time spent on review screen ($p = .003$)
  - Noticers spent an average of 130 seconds on review screen, mean was 40 seconds for non-noticers

- Anomaly type ($p = .02$)
  - Undervotes more likely to be noticed than flipped votes (61% vs. 39%)
### Results: Anomaly Detection

1. **Self-reported care in checking review screen**  
   
   - Self-reported care in checking review screen ($p = .04$)

<table>
<thead>
<tr>
<th></th>
<th>Not at all</th>
<th>Somewhat Carefully</th>
<th>Very Carefully</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detected</td>
<td>0%</td>
<td>4%</td>
<td>47%</td>
</tr>
<tr>
<td>Did Not</td>
<td>6%</td>
<td>24%</td>
<td>19%</td>
</tr>
<tr>
<td>Total</td>
<td>6%</td>
<td>28%</td>
<td>66%</td>
</tr>
</tbody>
</table>

2. **Information condition (marginal, $p = .10$)**

<table>
<thead>
<tr>
<th></th>
<th>Undirected</th>
<th>Directed with roll-off</th>
<th>Fully Directed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection Rate</td>
<td>44%</td>
<td>42%</td>
<td>64%</td>
</tr>
</tbody>
</table>
Results: Anomaly Detection

- Suggestive, but not statistically significant
  - The number of anomalies ($p = .10$)
    - Some evidence that 1 anomaly is harder than 2 or 8
  - The location of anomalies ($p = .10$)
    - Some tendency for up-ballot anomalies to be noticed more

- Non-significant factors
  - Age, education, computer experience, news following, personality variables
Results: Errors (Effectiveness)

No system was significantly more effective than the others.
Results: Error Types

![Error Rate Chart]

- Overvote Errors: 0.2 ± 1 SEM
- Undervote Errors: 0.4 ± 1 SEM
- Wrong Choice Errors: 1.6 ± 1 SEM
- Extra Vote Errors: 0.6 ± 1 SEM

Error Type

Mean Error Rate (%) ± 1 SEM
Results: True Errors vs. Residual Vote

- At the aggregate level agreement was moderate
- However, agreement was poor at the level of individuals
- For DREs:  
  \[ r(32) = .30, p = .10 \]
- For others:  
  \[ r(32) = .02, p = .89 \]
Results: Efficiency

- The DRE was consistently slower than the non-DRE voting technologies.

- Noticing of the anomalies was not a significant factor in overall DRE completion times.
Results: Satisfaction, Non-noticers

- Those who did not notice an anomaly preferred the DRE
  - Despite no clear performance advantages
  - Replicates previous findings
Results: Satisfaction, Noticers

However, if an anomaly was noticed, voter preference was mixed.
Despite our GUI improvements, only 50% of voters noticed up to 8 anomalies on their DRE review screen

- While this is an improvement over Everett (2007), half of the voters are still not noticing anomalies
- Data suggest that the improvement is mostly in detecting anomalous undervotes (orange highlighting helps!)
  - But vote flipping is still largely invisible
- This suggests that simple GUI improvement may not be enough to drastically improve anomaly detection
Discussion

- **VVPATs**
  - If voters are not checking review screens, how likely are they to check an external paper record?

- **Residual vote rate**
  - The relationship between the residual vote rate and the true error rate may not be straightforward
  - May be dangerous to simply assume correspondence

- **Subjective vs. objective performance**
  - In general, no strong association between preference and performance
  - However, voters who noticed the anomalies were less satisfied with the DRE