An Examination of the Auditability of Voter Verified Paper Audit Trail (VVPAT) Ballots

Stephen N. Goggin and Michael D. Byrne
Department of Psychology
Rice University, MS-25
Houston, TX 77005 USA
{goggin, byrne}@rice.edu

ABSTRACT

With heightened concerns over the security of electronic voting machines, 37 states now require some form of voter verification mechanism that could then be used in a recount. This usually takes the form of a Voter-Verified Paper Audit Trail (VVPAT). Unfortunately, little is known about the usability of VVPATs for recount purposes. The current study examines the speed and accuracy of hand recounts of VVPATs. Participants counted completed VVPAT ballots which were based on those actually in use in DREs today. Two races from of a spool of 120 ballots were manually counted, which includes separating ballots from the spool and removing rejected ballots. This task was time-consuming and prone to high error rates, with only 57.5% of participants' counts providing the correct election results. Furthermore, ballot rejection rate interacted with the closeness of the race being counted; high rejection rate paired with a small margin of victory resulted in a particularly high error rate. This experiment raises serious concerns about the viability of conducting manual recounts or audits using current VVPAT technology.

INTRODUCTION

With heightened public concerns of the security of Direct Recording Electronic (DRE) voting machines, 37 states have opted to require physical copies of ballots. The most common implementation by vendors is known as the VVPAT, or Voter Verified Paper Audit Trail. These add-on devices are intended to provide a secondary record of the voter's intent, while also serving two important functions. VVPAT ballots are intended to be verified by the voter to ensure that they are accurate, and they are also intended to serve as a system for post-election audits or recounts.

The Help America Vote Act (HAVA) of 2002 stipulated that voting systems used in federal elections should "produce a permanent paper record with a manual audit capacity for such system" Title III, Sec. 301, a2Bi. The HAVA legislation also created the Election Assistance Commission (EAC), which further outlined the guidelines for VVPAT ballots in their Voluntary Voting System Guidelines (VVSG) in 2005. While these guidelines are not yet mandatory, they are often met by vendors in terms of VVPAT construction and procedures.

Vendors have generally responded to the demand for the implementation of VVPAT systems to accompany DRE voting machines. Most vendors simply produce devices that can be attached to the voting machines already in use. These add-on devices often consist of a simple thermal paper printer with spools of receipt-width thermal paper serving as the ballots. This spooling system allows the VVPAT mechanism to be a rather simple mechanical device that requires no physical voter interaction. However, the implementation of VVPATs has presented many practical problems. Because of the added printer device, voting machines now require much more attention from poll workers to ensure that all parts are operating correctly. These thermal paper spools have already demonstrated numerous problems for election administration.

In fact, the history of VVPAT implementation has been wrought with problems. With the May 2006 Primary Elections in Cuyahoga County, Ohio, numerous problems involving the use of VVPATs were discovered. The Cuyahoga County Commissioners brought the Election Science Institute (ESI) to examine all aspects of their voting process. The ESI report details numerous findings regarding the VVPATs. For example, 10% of VVPAT spools were destroyed, blank, illegible, or missing; 19% of the VVPAT tapes indicated a discrepancy between the tallies of ballots (ESI, 2006). The ESI project team also offers recommendations for the future, including developing a specific audit procedure by the Board of Elections of Cuyahoga County, and the development of a "recount fixture" that can be used to count VVPAT ballots

(p.100). The ESI report highlighted the main weakness of the VVPAT ballots—they require extra attention from not only the voters, who are expected to verify their ballots (Cohen, 2005), but from the election administration that has to ensure the printers are also adequately working at all times during the election. With VVPATs implemented, election officials are now forced to ensure the proper functionality of the systems throughout election day, ensure adequate supplies of paper are available, and maintain physical custody of thermal paper spools from each machine, which are very sensitive to heat. The added complexity and amount of work for election officials increases dramatically, allowing more room for machine malfunction due to human error from poll workers.

The ESI report also demonstrates that handling VVPATs after an election can be difficult. While the report discusses numerous problems with the ballots themselves, it is unknown how well these ballots can be accurately tallied. Most previous literature regarding that of the usability of VVPAT ballots (e.g., Cohen, 2005) deals with the usability for the voter, not from the perspective of an auditing election official. Previous literature regarding the hand-counting of ballots never discusses the possibility of human error in counts; rather, this hand-count is trusted as a gold standard against which other voting methodologies "error" is measured (Ansolabehere & Reeves, 2004).

In April 2006, the Georgia Secretary of State released a report describing a pilot project in three counties to assess the feasibility of VVPATs being implemented statewide. This report detailed potential costs and reported timing data for the three-person audit teams that undertook the pilot project. The group found that the average time used to audit a single VVPAT ballot was 11 minutes. No error data were recorded regarding the audit because the true vote count could not be known. Most strikingly, the report predicts a cost of $3.01 per vote cast to conduct an audit. Therefore, a full recount of VVPAT ballots in Cobb County, Georgia, where 179,652 votes were cast in the November 2006 General Election would cost the county $540,753.

Our research examines the auditability of VVPAT ballots using the three metrics of efficiency, effectiveness, and satisfaction proposed by the National Institute of Standards and Technology (NIST) in their 2004 Special Report (Laskowski et al., 2004). Although auditability is a measure of accuracy, and therefore effectiveness, it is also important to know with what speed an audit system can operate under, and what satisfaction or strain it places on those conducting the audit. This paper examines a best-case scenario from the voter perspective; that is, it assumes all votes are cast accurately and verified to be correct. We examined with what accuracy people can manually audit or recount VVPAT ballots from an election. We hoped to determine whether the properties of VVPAT ballots makes them an appropriate medium for auditing. If VVPAT ballots are meant to be a double check on the integrity of computerized voting systems, then the auditing system should not be readily prone to error.

METHOD

Participants

Forty Rice University undergraduates participated, for which they were compensated with credit towards a course requirement. There were 14 male and 26 female participants, with a mean age of 19 years. All participants had normal or corrected to normal vision, and were fluent in English.

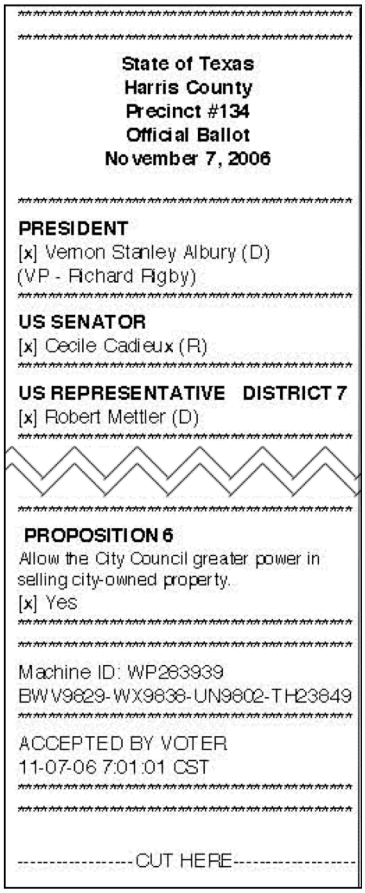
Materials

The participants counted specially made spools of VVPAT ballots, which were constructed by strictly following the 2005 VVSG standards in section 7.9.6 regarding VVPAT usability (pp.143–144), and similar to those that would be stored in a DRE machine with VVPAT capability after an election. As the election progresses, VVPAT ballots are wound in reverse order onto a secondary spool, where they are stored until they are counted. Each spool was composed of 120 ballots, which were randomly generated by a computer program.

Each ballot contained twenty-seven political races and propositions, with the candidates and propositions having fictional names and properties; these ballots were identical to those originally used by Everett, Byrne and Greene (2006). All ballots contained time stamps and identifying information similar to an actual ballot (See Figure 1). Each ballot also contained a notation at the bottom, signaling whether it had been "Accepted" or "Rejected" by the voter, in accordance with requirement 7.9.2 in the VVSG (p. 137). The spool was placed on a stand, or a "recount fixture" that allowed participants to unroll the ballots and separate them as they counted them (See Figure 2).

To make the ballots closely match those that might be cast in a real election, the ballot roll-off, or rate of abstention as a function of ballot length, was made higher for the races further down the ballot. Based on data from Nichols and Strizek (1995) regarding roll-off on electronic voting machines, the abstention rate for

the top race was set at 9%, while it was 15% for the lower race.


State of Texas
Harris County
Precinct #134
Official Ballot
November 7, 2006

PRESIDENT
[x] Vernon Stanley Albury (D)
(VP - Richard Rigby)

US SENATOR
[x] Cecile Cadieux (R)

US REPRESENTATIVE DISTRICT 7
[x] Robert Mettler (D)

PROPOSITION 6
Allow the City Council greater power in selling city-owned property.
[x] Yes

Machine ID: WP283939
BWV9829-WX9838-UN9802-TH23849

ACCEPTED BY VOTER
11-07-06 7:01:01 CST

----------CUT HERE----------

Figure 1. Partial Sample Ballot

Procedure

Participants first filled out a background demographic survey, after which they were handed two sheets of detailed instructions for the VVPAT counting task. The instructions were designed to be as clear and concise as possible, providing a step-by-step process the participants should use to count the ballots. Participants were told to first physically separate the ballots from the spool on the recount fixture using scissors, setting aside all "rejected" ballots. Next, they were told to count the ballots for the specific race, tallying the vote and providing a final total on their tally sheet. After participants read the directions, the experimenter verbally reiterated key instructions, and participants were allowed to ask questions about the process. To be completely sure participants understood the task, the experimenter also showed sample ballots to the participants, pointing out the relevant notations and layout of the ballot. After the participant indicated confidence in their ability to complete the task, they were allowed to begin counting. Participants were allowed to keep the instruction sheets with them during the task and were allowed to ask any questions if they were unsure of the process during the task.

Figure 2. Recount Fixture

To begin the task, participants were handed a sheet on which they were told to tally the votes of one of the races and record an official total at the end of the counting session. They then counted the request race, and upon completion of that, they were then prompted to count a second race using the same ballots they had just counted, although this time, they did not have to cut them from the spool. After they finished counting this race in the same manner, participants filled out an evaluation questionnaire about the counting process.

The procedure was strictly derived from the actual counting procedures for VVPATs recommended by the VVSG and thus used by many election officials across the United States. Our procedure differs in one critical way: election audits or recounts often use teams of two or three counters to increase accuracy. This is an important safeguard; we did not do redundant counting because we wanted to examine the actual base rate of error, that is, we wanted to determine the extent of the problem against which the redundancy guards.

Design

A mixed design with one between-subjects variable and one within-subjects variable was used. The between-subjects variable was the rate of "rejected" ballots in the spool. In one condition, 8 ballots out of 120 (6.6%) were rejected, while only 4 ballots (3.3%) were rejected in the other condition. The within-subjects variable was

the closeness of the chosen races on the ballot. In the lopsided condition, the winning candidate won with a margin of roughly 30% of the total vote, while in the close condition, the winning candidate won with a margin of roughly 5% of the total vote.

The three dependent variables were measured in this study were errors, counting time, and satisfaction. We also asked several open-ended questions allowing for participant comments on aspects of the counting process. Obviously, the most important metric for any recount or audit system is its accuracy. Errors were studied as both absolute values and signed differences from the correct ballot counts for each candidate. Also, the obtained difference from the true margin of victory was measured in both absolute and signed terms to determine the effect on the outcome of each race. Time was measured in seconds from when the participants were handed the tally sheet to when they completed the counting process for one race. Satisfaction was measured at the end of the counting process using a common standardized instrument, the System Usability Scale (SUS) (Brooke, 1996). Finally, participants were asked about their confidence in the accuracy of their counts, and responded to several questions about improving the counting process and the design of the ballots.

RESULTS

Effectiveness

Error was measured at the individual candidate level, so 160 error percentage observations were recorded for the 40 participants, since each participant counted two two-candidate races. The range was from -17% to 19% (M = 0.17%, SD = 3.3%). As clearly shown in Figure 3, the error rate was especially high for lopsided races in the high reject rate condition. An ANOVA revealed that there was a significant interaction of the closeness of a race and the rejected rate on the signed percentage error rate, $F(1, 36) = 5.19$, $p = .029$. No other main effects or interactions were significant.

In order to quantify the effect of error better without the signed values canceling, the absolute value of these error scores were taken. While an ANOVA revealed no significant differences between groups, the total error averaged across all counts is much larger (M = 1.3%, SD = 3.0%) than the signed rate calculated above.

When the error variable was defined as the margin of victory for the winning candidate and the absolute percentage discrepancy from this margin, the error rate becomes large. Of the 80 counted races, correct counts were only provided for 57.5% of the margin scores, and the error average error rate was quite large (M = 6.2%, SD = 11.0%). Races with close margins of victory had less error (M = 3.5%, SD = 6.3%) than those with lopsided victories (M = 8.9%, SD = 1.4%); this effect was statistically reliable, $F(1, 36) = 5.16$, $p = .029$.

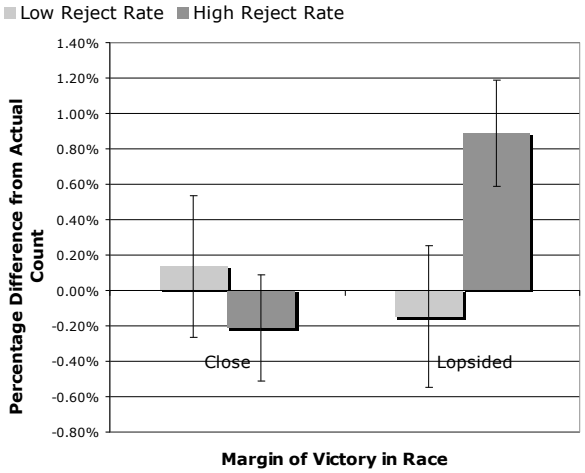


Figure 3. Overcount and Undercount Error Rates by Candidate Vote Totals ± 1 SEM

Efficiency

With any proposed audit system, it is also important to consider the time it would take for an audit to be conducted. A mixed 2x2x2 ANOVA (including an order variable: first race counted vs. second race counted) revealed no significant differences in time spent counting based on the rate of rejected ballots or closeness of the races. An order effect was clearly demonstrated, $F(1, 36) = 240.22$, $p < .001$, however this is explained by the task of cutting the ballots in the first count, whereas no cutting was required in the second count. Because the tasks of cutting and counting the ballots were completed concurrently for the first count done by each participant, no estimation of performance improvement between the first and second counts could be calculated.

While no differences between conditions were observed, the timing data yield clear insight on the lengthy nature of the task. In the first task, in which participants had to manually separate the ballot from the spool, the participants took much longer (M = 25.3 min, SD = 6 min), while when the participants only had to count the separated ballots, there times were much shorter (M = 11.9 min, SD = 4.0 min).

If we assume standard counting procedures for most counties, with ballots would counted by three

independent counters as done in the Georgia pilot project, we can extrapolate an estimate of the time involved in a audit involving an entire county's ballots. Because the ballots would still have to be separated by one worker, we can use the time for each task to represent the average times for each person doing the count. Therefore, for 120 ballots, we estimate with 95% confidence that it would take between 0.74 and 0.85 hours of labor *per race* for an audit of 120 ballots to be completed.

Satisfaction and Subjective Measures

Also of interest to potential auditors or counters is the ease of use and satisfaction of a system used in such a capacity. With the SUS ranging from 0-100 with 100 being most satisfied with a system, the mean score was 63.2, with a standard deviation of 13.6. The ANOVA revealed that SUS scores did not significantly differ based on the reject rate, $F(1, 36) = 0.15$, $p = .71$, or the closeness of races, $F(1, 36) = 0.08$, $p = .78$. Unsurprisingly, this reflects a low level of subjective satisfaction.

Participants were also asked to rate the confidence in his/her counts on a 1–5 Likert scale, with 5 being completely confident in the accuracy of the count. The confidence levels reported indicate a lack of certainty in the counts (M = 3.98, SD = 0.69). More interestingly, confidence was not significantly correlated with any of the measures of error; the largest correlation was with the absolute error scores and this correlation was actually negative (meaning more confident participants actually made slightly more errors), $r(38) = -.15$, $p = .37$.

Participants were also asked to write complaints or comments about ways in which the process of counting could be improved. Of the 40 participants, 65% commented that the process was subject to natural human error due to the tedious or repetitive nature of the task, and 25% complained about the process of separating the individual ballots from the roll. Most of the comments were directed at the design of the ballot and the problems it creates in the audit process. 20% suggested reordering or numbering the ballot, 22.5% suggested a clearer font or color be used on the ballot, 27.5% suggested rearranging the ballot into separate receipts for each race, while another 40% complained about the length of the 27 race ballot.

DISCUSSION

Because the VVPAT is intended as an audit device to ensure the accuracy of elections, errors of any level in their counts should raise alarm. Hand counts have usually been considered the standard against which other technology's accuracy is gauged; however, human counting performance is not perfect. As the participants commented, the tedious and repetitive nature of the test can lead to oversight of a step in the process and therefore an error. While having a second person recount the ballots would reduce the error rate, the high error rates from single counters suggest that an error-free count is unlikely even if counted twice.

Furthermore, while undergraduate students from a highly selective university likely differ from typical (generally older) counters in completing audits or recounts, the most regular finding in the substantial literature in the study of human aging (e.g., Salthouse, 1991) is that older adults perform tasks more slowly than younger adults. Thus, our findings likely represent a best-case scenario in terms of efficiency. However, there is no strong data that suggest substantial differences in effectiveness or satisfaction on the basis of age.

The interaction between the margin of victory in the races and the number of rejected ballots in the spools provides evidence for participant sensitivity to the content of the count. One would naturally expect the ballot spools with the higher rate of rejected ballots to produce more errors; however, because this effect is only present when the victory in the race is lopsided, this suggests that participants were less attentive to the rejected notation when they detected that the race was not close. This discovery suggests that biases in counts, even unintentional ones, may result from counters' beliefs about the margin of victory. Whether this effect can be generated by prior beliefs (e.g., races in districts where one party has a historical advantage) alone and not by observing the margin as in our experiment is unknown.

In addition, because of the needed accuracy in recounts, and the tedious nature of unspooling, separating, and counting VVPAT ballots, a large recount would be a laborious and costly process. If a recount by three workers of one race on 120 ballots takes 0.74 to 0.85 work hours, then a complete recount of a large county, such as Cuyahoga County, Ohio, with 673,740 voters in the 2004 Presidential election, would take between 4155 and 4772 hours of labor to recount only one race. That is slightly more than two person-*years* assuming a standard 2000-hour work year. This laborious recounting process strongly suggests that using VVPATs as a check on DRE systems may not be

practical.

Because baseline measures of effectiveness, efficiency, and satisfaction have not been recorded for manual recounts of other ballot types, we cannot conclude that VVPAT manual recounts are necessarily different than manual recounts of other types of ballots. For example, manual recounts of "hanging chad" punch card ballots are notoriously slow. However, features of the VVPAT ballots which were criticized by participants, such as the low quality of text produced by thermal printers and difficulties in dealing with spools of lengthy ballots, make it clear that this method can be improved.

Simply because VVPAT ballots are intended to be readable by humans does not mean that more automated mechanisms could not be used in a recount of these ballots to help solve problems with efficiency and error. Because VVPAT ballots are produced with a standardized character layout, they could be adapted to be read by optical scanning technology. It may be that the most important property of VVPATs is not the hand count, but the fact that voters can actually verify their selections on a record that is physically separate from the DRE. This implies that any count of the VVPATs, be it machine or human, would be more trustworthy than the unobservable bits in a DRE (assuming, of course, that the VVPAT counting hardware is trustworthy). However, greater flexibility in conducting audits or recounts of the ballots may be beneficial for local election officials.

While VVPAT ballots are intended to check voting system security by having another copy of a voter's ballot stored physically separate from the electronic copy, if these ballots cannot be easily counted without error, they do not fulfill their role as a reliable audit system. While some safeguards, such as multiple counters, are often used to increase the accuracy of an audit or recount, a well-designed audit system should be easily counted without error, minimizing (though not necessarily eliminating) the need for costly safeguards and repetition. Of course, VVPATs may have value for other reasons, but whether those justify the cost is unclear.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under grant #CNS-0524211 (the ACCURATE center). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of the NSF, the U.S. Government, or any other organization.

REFERENCES

Ansolabehere, S. & Reeves, A. (2004). Using Recounts to Measure the Accuracy of Vote Tabulations: Evidence from New Hampshire Elections 1946-2002. <http://vote.caltech.edu/media/documents/wps/vtp_wp11.pdf>

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.

Cohen, S.B. (2005). Auditing Technology for Electronic Voting Machines. Unpublished master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

Election Science Institute. (2006). DRE Analysis for May 2006 Primary Cuyahoga County, Ohio. San Fransisco, CA. <http://bocc.cuyahogacounty.us/GSC/pdf/esi_cuyahoga_final.pdf>

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

Georgia Secretary of State, Elections Division. (2007). Voter Verified Paper Audit Trail: Pilot Project Report, SB500 2006 Georgia Accuracy in Elections Act. <http://www.sos.state.ga.us/elections/VVPATreport.pdf>

Help America Vote Act of 2002, H.R. 3295, 107th Congress. (2002).

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.

Nichols, S.M. & Strizek, G.A. (1995). Electronic Voting Machines and Ballot Roll-Off. *American Politics Quarterly*. *23*(3), 300-318.

Salthouse, T. A. (1991). *Theoretical perspectives on cognitive aging*. Hillsdale, NJ: Erlbaum.