ELECTION LAW JOURNAL
Volume 11, Number 1, 2012
© Mary Ann Liebert, Inc.
DOI: 10.1089/elj.2010.0098

# Post-Election Auditing: Effects of Procedure and Ballot Type on Manual Counting Accuracy, Efficiency, and Auditor Satisfaction and Confidence

Stephen N. Goggin, Michael D. Byrne, and Juan E. Gilbert

## ABSTRACT

A task common to nearly all types of election auditing is that of human auditors physically counting ballots by hand. This task, fundamental to the goal of accuracy in an audit, can be a source of error. While somewhat basic in its nature, the process of counting can be strongly influenced by many procedural and legal factors. In the current study, we examine how specific group counting procedures and ballot types affect the accuracy, efficiency, and subjective judgments of usability of a post-election audit. These two procedures, quite different in their implementation and employed in real elections in two U.S. states, have built-in redundant checks and multiple tallies to help bolster accuracy; we found that even with this redundancy, errors are surprisingly frequent. Additionally, certain counting procedures are more efficient, as well as less variable in the amount of error they introduce into the audit process. We found that well-specified procedures, as well as division of labor amongst group counting members, help ensure more accurate and efficient ballot audits.

## INTRODUCTION

With the accuracy of elections in the United States raised as an issue of public concern following the 2000 presidential election, federal, state, and local jurisdictions have responded through a patchwork of new regulations, procedures, and recommendations. After the Help America Vote Act (HAVA) of 2002, many jurisdictions chose to upgrade their voting machines, as well as update many of the procedures used in the administration of elections. While there are many aspects of elections that remain ripe for study, including security, usability, reliability, legitimacy, as well as transparency, one could easily argue the underlying reason for all of these areas of concern is that of accuracy. For an election to be legitimate, it should accurately represent the intentions of all its voters.

In efforts to address the concern of accuracy, many jurisdictions have implemented post-election audit or recount procedures utilizing manual counting by election officials. These procedures, intended as a safeguard against computer and prior human error, are becoming commonplace across the United States, with 21 states requiring post-election manual audits of some magnitude, with an additional 17 states producing paper records capable of supporting a manual audit, although not explicitly requiring post-election audits (Verified Voting, 2009). However, despite these efforts to make election outcomes more accurate, all have operated under the untested assumption that a hand count of ballots could recover the truly accurate vote total. In practice, post-election audits or recount procedures encompass a large area of election administration.

Stephen N. Goggin is a Ph.D. student in political science at the University of California, Berkeley. Michael D. Byrne is an associate professor in the Department of Psychology at Rice University in Houston, Texas. Juan E. Gilbert is a professor in the School of Computing at Clemson University in Clemson, South Carolina.

From sampling of precincts, to chain of custody of ballots, to methods for resolving discrepancies, to reporting of results, many areas must be addressed carefully by election administrators.

Considerations of all of these areas of auditing have been discussed in numerous academic, non-profit, and governmental publications (e.g. Norden et al., 2007; Hall, 2008c; Duffy et al. 2009; New Hampshire Department of State, 2006; Atkeson et al., 2008). However, one procedure is common to all recount procedures: the manual counting of ballots by election officials, or auditors. While counting procedures have been described and discussed in publications such as the examples referenced above, few studies have experimentally examined the quality and usability of these procedures. Simple administrative or procedural changes, as well as the adoption of statutes that specifically govern procedures for auditing, might produce dramatic differences in the accuracy of hand audits. Moreover, the assumption that instituting a recount of ballots will monotonically increase the accuracy of an election is one that is highly relevant and should not be ignored in policy discussions. The accuracy, efficiency, and usability of procedures is something that can be empirically tested—there is no reason to simply rely on the assumptions and best guesses of varying election officials regarding proper auditing procedures.

This study addresses the simple question inherent in every manual recount and audit procedure: just how good are audit teams at counting? More precisely, independent of all the other areas of study in election auditing, how do ballots themselves and the counting procedure used affect the accuracy and efficiency of the count, as well as the satisfaction and confidence of the auditors? Should we use manual recounts as the "gold standard" by which accuracy of voting systems is measured? We characterize our dependent metrics for measuring these procedures, as well as the different types of ballot systems, and the different election auditing procedures used in the following sections.

## MEASURING USABILITY

While accuracy, as we have argued above, is generally considered to be the clear and crucial goal of elections, we must also measure audit procedures in other ways. That is, in the current study, we measure the usability of manual audits using three metrics, as recommended by the U.S. National Institute of Standards and Technology (Laskowski et al., 2004). These three metrics for usability include the objective metrics of effectiveness and efficiency, as well as the subjective metric of satisfaction. For an audit system to be usable, it therefore must be accurate (effective), efficient in time and cost, and should inspire confidence in the counts as well as leave the auditors satisfied with the procedure.

The task of auditing is a relatively simple procedure, however its usability is relatively unknown. As described below, the procedures used can be adequately described in a few sentences or less, and require little strain on auditor's memory for the key components of the task, as the labor is often divided among the audit team, or is cued directly from the procedure. Additionally, counting is a skill commonly learned early in life, and because of its ubiquitous use in everyday life, people are generally accustomed to the task. Even though the task is relatively simple, there remain many opportunities for error, both at the individual level—resulting from problems such as hearing impairment or low math skills—to those at the procedural level. When dealing with numerous complex ballots throughout lengthy audits, auditors can easily count ballots multiple times, skip over ballots, or misread a voter's intent.

There is little psychological research on group performance on such mundane tasks as counting. While there are numerous examples of complex tasks involving dyads and larger groups within psychology literature, the tasks in these studies are quite unlike a simple auditing task, and the studies focus on interaction between group members in less well-specified tasks, such as decision making or the learning of complex procedures (e.g., Webb, 1982; Watson, Michaelsen, and Sharp, 1991). A search of the applied psychology literature revealed no laboratory research utilizing groups for similar clerical or counting tasks. For our purposes, the task given to audit teams is essentially a vigilance task, demanding concentration and carefulness, requiring little planning and few, if any, unscripted interactions with other group members.

## GROUP COUNTING PROCEDURES

In the current study, we utilized two prototypical group counting procedures, one with a four-person audit team, and one with a three-person audit team, to count several different types of ballots. The two procedures, described in detail below, are commonly called the "read-and-mark" and the "sort-and-stack" method. Because of the daunting patchwork of regulations and recommendations of procedure types, we do not attempt to exhaustively address all types of group counting. However, we chose two methods that represent two distinct counting styles, as well as different levels of redundancy. Additionally, both of these methods are currently employed in jurisdictions throughout the country with minor modifications.

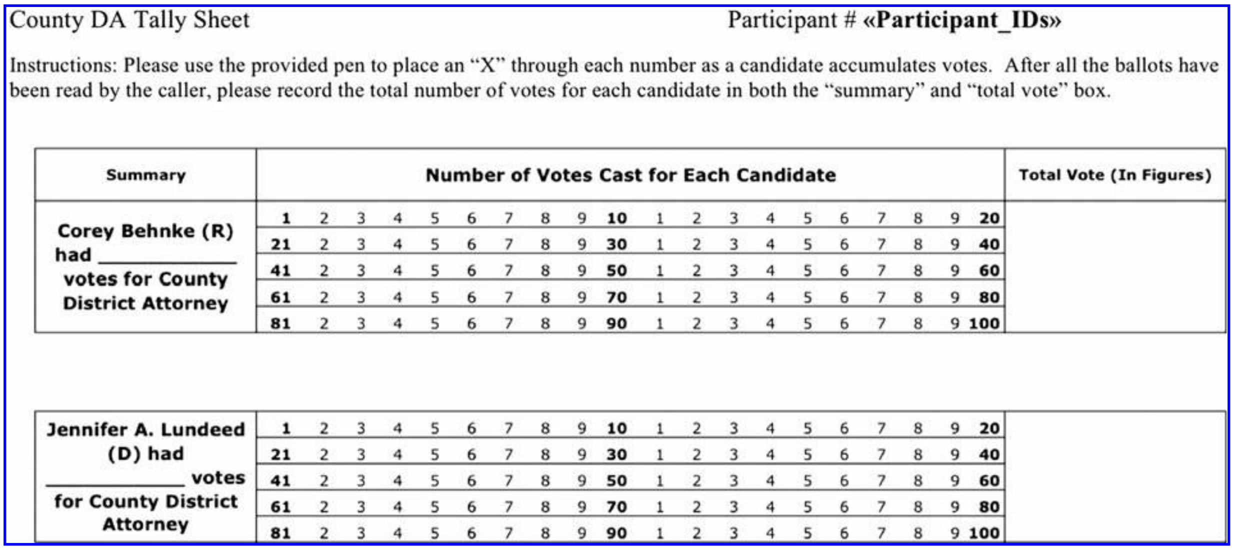
These ballot counting procedures must handle several issues that arise when counting ballots. A ballot that auditors encounter may have valid selections for the candidates in the political race of interest, or they may be under-voted or over-voted. An under-voted ballot item means that the voter abstained, or did not choose an option in that race, while an over-voted ballot item means that more candidates than allowed for a specific race were selected, which renders the ballot invalid. In real elections, auditors may encounter ballots with more ambiguity and problems, such as write-in candidates; for our purposes, all ballots fall into one of three categories: valid, an under-vote, or invalid because of over-vote.

The first procedure, sometimes termed a read-and-mark method (Stevens, 2007), utilizes four election officials, and can be thought of as a top-to-bottom audit procedure. That is, the auditors count the ballots sequentially as they are taken from the top of the unsorted stack of ballots. Additionally, it can be thought of as a concurrent task because all election auditors are involved in the procedure at the same time.

The first member, the caller, "speaks aloud the choice on the ballot for the race being tallied." The second team member, the witness, "observes each ballot to ensure that the spoken vote corresponded to what was on the ballot and also collates ballots in cross-stacks of ten ballots." The final two members of the audit team are both talliers, who "record the tally by crossing out numbers on a tally sheet to keep track of the vote tally." Additionally, the talliers "announce the tally at each multiple of ten…so that they can roll-back the tally if the two talliers get out of sync" (Hall, 2008b, p. 3). This procedure also utilizes a specialized tally sheet with pre-numbered blocks, which allows the tallier to simply mark each number as it is called, rather than creating individual tally marks.

This procedure was used in San Mateo County, CA in their 2006 and 2007 elections, and similar procedures are described in use elsewhere (Hall, 2008a; Hall 2008b; Alvarez, Katz, and Hill, 2005). This method (see Figure 1) is specified in the county election procedures for multiple California counties, and is described in CA Election Code 15102, although only specifically for counting vote-by-mail ballots. (Hall, 2008b, p. 3).

The second procedure, sort-and-stack, like the read-and-mark procedure, only counts one race at



County DA Tally Sheet Participant # **«Participant_IDs»**

Instructions: Please use the provided pen to place an "X" through each number as a candidate accumulates votes. After all the ballots have been read by the caller, please record the total number of votes for each candidate in both the "summary" and "total vote" box.

| Summary | Number of Votes Cast for Each Candidate | | | | | | | | | | | | | | | | | | | | Total Vote (In Figures) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Corey Behnke (R) had ____________ votes for County District Attorney | **1** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **10** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **20** | |
| | **21** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **30** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **40** | |
| | **41** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **50** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **60** | |
| | **61** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **70** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **80** | |
| | **81** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **90** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **100** | |
| Jennifer A. Lundeed (D) had ____________ votes for County District Attorney | **1** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **10** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **20** | |
| | **21** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **30** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **40** | |
| | **41** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **50** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **60** | |
| | **61** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **70** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **80** | |
| | **81** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **90** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **100** | |

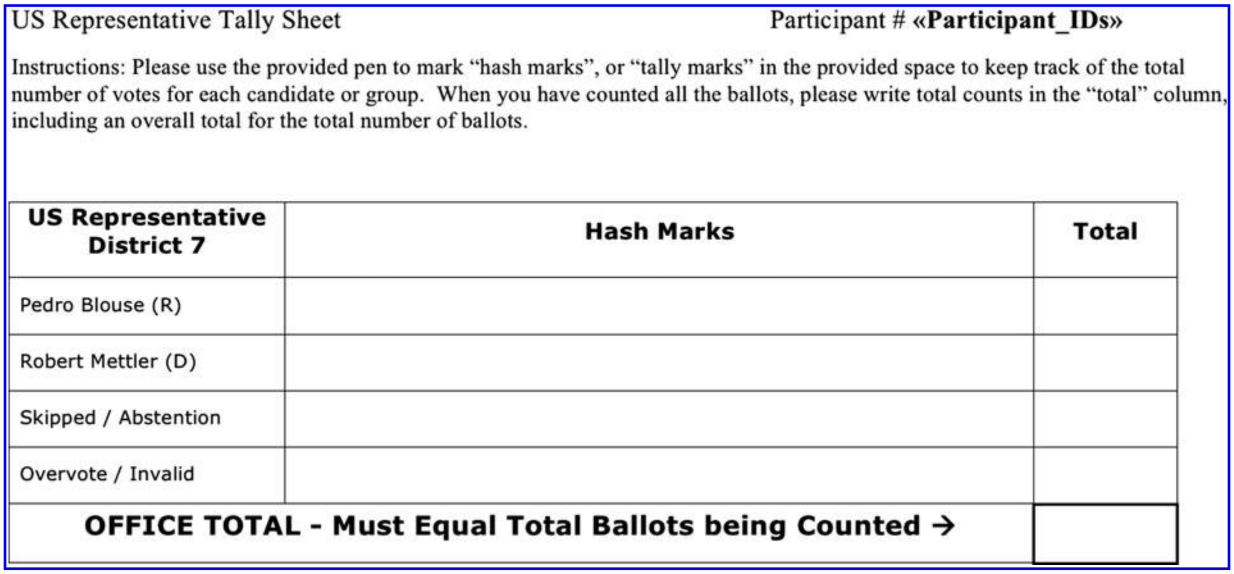
**FIG. 1.** Read-and-mark tally sheet.

a time. In our study, we employed a three person audit team, Unlike the read-and-mark procedure, however, the roles and labor needed for the counting task is not divided among the team members. As the name indicates, the team members first sort the ballots into several piles, based on the choice in the race being audited. For instance, a race between Candidate A and Candidate B results in four piles: one for each candidate, one for under-voted ballots, and one for over-voted ballots. As these ballots are being sorted into their respective stacks, they are placed in cross-stacks of ten ballots. Once the piles are separated, each auditor counts each stack once, and the tallies are compared to check for accuracy. All three members of the counting team have their own tally sheets. An example tally sheet is shown in the New Hampshire Elections Procedure Manual (EPM) (New Hampshire Department of State, 2006, p. 157), and unlike the read-and-mark tally sheet, auditors are provided with blank space in which to make their own tally marks as the votes are counted.

The sort-and-stack procedure (Figure 2) is described in detail in the New Hampshire EPM (New Hampshire Department of State, 2006, p. 149–152). This method is described in numerous other publications, including presentations by the New Hampshire Assistant Secretary of State (Stevens, 2007) and literature from the Election Defense Alliance and the NH Fair Elections Committee (Tobi, 2010). Additionally, Halvorson and Wolff (2007) note a very similar procedure's use in Minnesota's 2006 post-election audit. As noted in the EPM (p. 149), the outlined sort-and-stack procedure is simply a suggested one, and the moderator in each jurisdiction in NH has the authorization to choose the system of counting and supervise it.

The sort-and-stack procedure described above contains two slight differences from that described in the EPM. First, the cross-stacks of ten were added to this procedure to make it more comparable to the read-and-mark procedure, providing a intermediary check of accuracy, allowing for auditors to more easily roll back the tally if counts disagree. Second, we had participants count each ballot stack separately to make the procedure more redundant, unlike the read-and-mark procedure, in which each ballot is only handled once. This difference from the read-and-mark procedure helps establish a test of whether the interaction of team members and division of labor is beneficial, or whether a more compartmentalized, redundant counting procedure performs better.

As noted by Olson (2009) in his description of post-election audits in many counties in Ohio conducted after the November 2008 election, many jurisdictions used similar read-and-mark methods, while some appeared to use methods more similar to the sort-and-stack method. He also details numerous problems with ballots and how discrepancies were resolved, often on a case-by-case basis.



US Representative Tally Sheet Participant # **«Participant_IDs»**

Instructions: Please use the provided pen to mark "hash marks", or "tally marks" in the provided space to keep track of the total number of votes for each candidate or group. When you have counted all the ballots, please write total counts in the "total" column, including an overall total for the total number of ballots.

| **US Representative District 7** | **Hash Marks** | **Total** |
|---|---|---|
| Pedro Blouse (R) | | |
| Robert Mettler (D) | | |
| Skipped / Abstention | | |
| Overvote / Invalid | | |
| **OFFICE TOTAL - Must Equal Total Ballots being Counted →** | | |

**FIG. 2.** Sort-and-stack tally sheet.

Depending on the type of ballots used and the condition of the records, as well as the discretion of the local election official, there are numerous ways of implementing a post-election audit. Additionally, in efforts to improve administration of elections and post-election audits, many interested groups have provided recommendations of specific methods by which to conduct post-election audits, including the counting of ballots as well as the other aspects of auditing from start to finish. (Duffy et al., 2009; ElectionAudits.org, 2008; Norden et al., 2007; Ohio Joint Audit Working Group, 2008) All these publications, while providing valuable tools and recommendations for procedural improvements, offer varying and sometimes vague methods for counting ballots, which has led to many permutations and adaptations of the two fundamental methods: read-and-mark and sort-and-stack, on which we focus. We analyzed a best-case scenario, with a well-specified prototypical procedure and unambiguous ballots for each method.

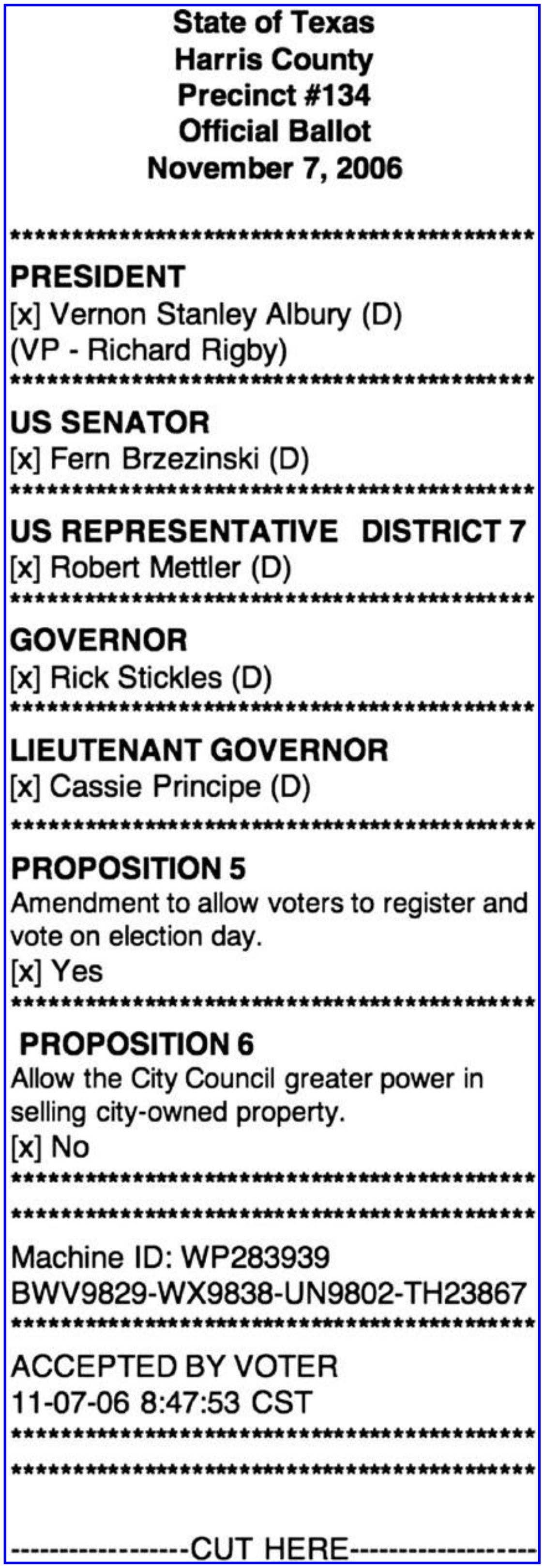
## BALLOT SYSTEMS

Different electoral jurisdictions also utilize different voting methods and voting technologies; local election officials often have near complete control over the ballot systems and voting methods in their area. While these voting methods have been studied with regard to their usability for voters, election officials must also interact with the machines or methods, making the study of their usability for election officials also of concern. Of great concern to election auditors, these different methods produce fundamentally different ballots. In the November 2008 election, 55.9% of United States counties utilized optical scan voting systems, while 34.3% of counties used Direct Recording Electronic (DRE) systems (Election Data Services, 2008). Lever machines, punch-cards, as well as hand-counted paper ballots were each used in less than 2% of counties. Because nearly all ballots used were either optical scan or from DRE machines, we examine how three different types of ballots, all coming from optical scan or DRE systems, perform in hand count audit procedures. We examine a prototypical thermal paper Voter Verifiable Paper Audit Trail (VVPAT) ballot (Figure 3), a legal-sized (8.5 inch by 14 inch) VVPAT ballot printed from the Prime III (Cross et al., 2007; McMillian et al., 2007) (Figure 4), as well as a legal-sized optical scan “bubble” ballot (Figure 5).

With security concerns over DRE voting machines, many states have required and utilize additional modules which produce printouts of each voter’s ballot. These Independent Voter-Verifiable Records (IVVR), as recommended by the Election Assistance Commission, Voluntary Voting System Guidelines (VVSG) (2007, 4.4.1.A2–3), provide an independent, physical copy of every ballot cast. With DRE machines, this entails some form of physical printout onto paper. These IVVRs, as noted in Goggin and Byrne (2007), fulfill a dual purpose. First, they serve as a review mechanism for the voter, allowing them to check their choices before casting their ballot. Second, they serve as a physical audit trail for post-election auditing, creating an indelible record of voter intent. Because IVVRs only fulfill these two purposes, they should be designed with clarity for both voters and election auditors.

However, because IVVRs were introduced post hoc for many DRE systems, they are often implemented as a printer module that attaches to the side of the voting machine. This style of IVVR (Figure 3), which we term a thermal paper VVPAT (although VVPATs can be printed in other ways), consists of a thermal paper spool behind a sheet of glass, and is described in detail in Goggin and Byrne (2007) and elsewhere. Election officials must then count the spools from the thermal printer modules, with some jurisdictions separating the ballots for counting, while others leaving the spool intact for the counting procedure. Because the VVPAT spools are a continuous sheet of paper, any spoiled ballot that a voter rejects as inaccurate appears alongside valid ballots, and cannot be removed during the election. Because of this, election auditors must be vigilant of voided ballots as they count and not include them in the vote tally. These voided ballots are different than the under-voted and over-voted ballots we discussed above; they are merely voided attempts at casting a ballot that a voter made, only to cast a valid ballot afterwards.

In order to provide a comparison for the thermal paper VVPATs, we also examined a VVPAT created by the Prime III system (Cross et al., 2007; McMillian et al., 2007), which is a prototype DRE system designed for multi-modal interaction for accessibility, as well as security (Figure 4). Unlike the thermal paper VVPAT, the Prime III VVPAT uses a standard



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

**********************************************

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

**********************************************

US SENATOR
[x] Fern Brzezinski (D)

**********************************************

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

**********************************************

GOVERNOR
[x] Rick Stickles (D)

**********************************************

LIEUTENANT GOVERNOR
[x] Cassie Principe (D)

**********************************************

PROPOSITION 5
Amendment to allow voters to register and vote on election day.
[x] Yes

**********************************************

PROPOSITION 6
Allow the City Council greater power in selling city-owned property.
[x] No

**********************************************

**********************************************

Machine ID: WP283939
BWV9829-WX9838-UN9802-TH23867

**********************************************

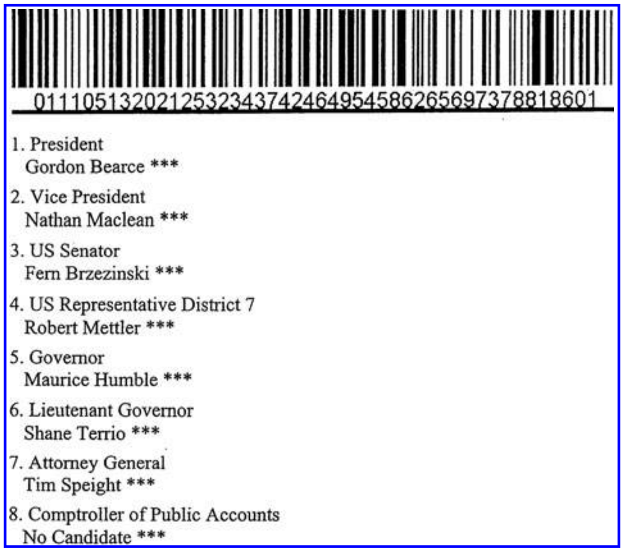
ACCEPTED BY VOTER
11-07-06 8:47:53 CST

**********************************************

**********************************************

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

**FIG. 3.** Thermal VVPAT example.



0111051320212532343742464954586265697378818601

1. President
Gordon Bearce ***

2. Vice President
Nathan Maclean ***

3. US Senator
Fern Brzezinski ***

4. US Representative District 7
Robert Mettler ***

5. Governor
Maurice Humble ***

6. Lieutenant Governor
Shane Terrio ***

7. Attorney General
Tim Speight ***

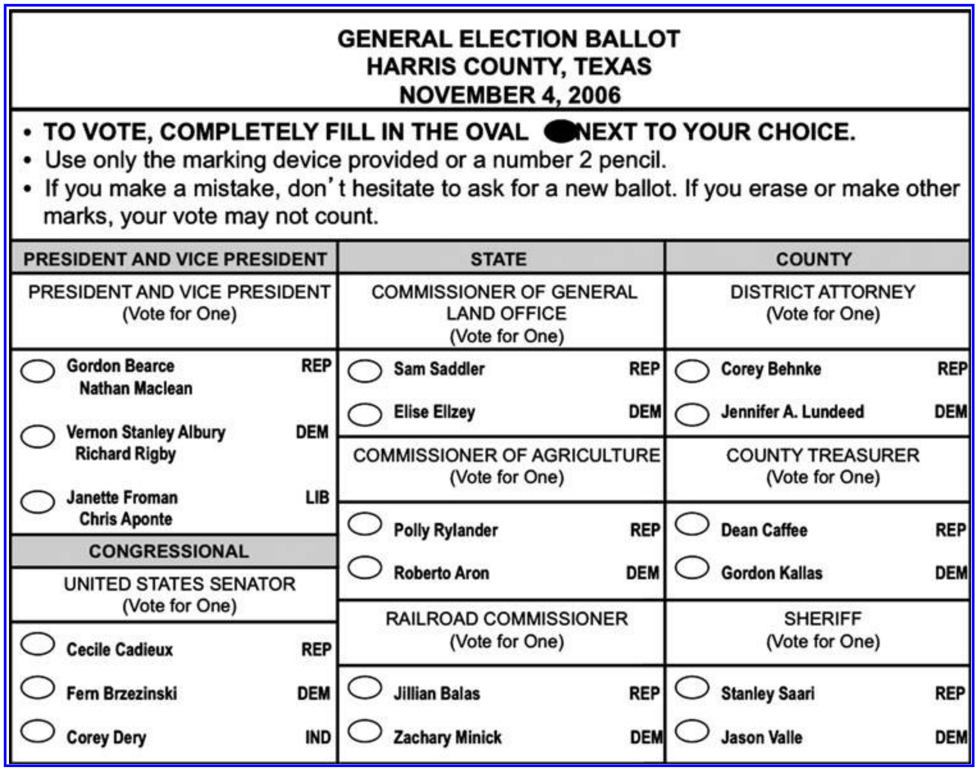
8. Comptroller of Public Accounts
No Candidate ***

**FIG. 4.** Prime III VVPAT example.

office printer to create the VVPATs, with each ballot as a separate legal-sized sheet. Like the thermal paper VVPAT, the Prime III VVPAT provides a record of only the balloted choice in each race. That is, each ballot only displays the candidate or response for which the voter cast their ballot, along with a heading indicating the race.

The third voting technology and ballot style we examined is an optical scan ballot system (Figure 5). Because the optical scan ballot is already a machine-independent record, it is an IVVR, with no separate record needed. Unlike the VVPAT and other ballot records generated by DRE systems, an optical scan ballot is the paper ballot that the voter physically interacts with and marks. Once the voter finishes marking their selections, depending on the election procedure used, the ballot is either scanned in front of the voter, or is safely deposited in a ballot box for scanning at the close of polls.

Because the voter interacts physically with the ballot, marking their choices with a pencil or other marking device, there is ample opportunity for a voter to create ambiguous marks, either by stray marks, smudges, or incorrectly filled in portions of the ballot. Unless the ballot is scanned in front of the voter (and even then, it may not be noticed), these ambiguous marks may lead to a voter's ballot being disputed and not counted. While the interpretation of these ballots is of utmost concern in real election auditing, as has been witnessed with the 2008 Minnesota Senate recount (examples of



GENERAL ELECTION BALLOT
HARRIS COUNTY, TEXAS
NOVEMBER 4, 2006

- TO VOTE, COMPLETELY FILL IN THE OVAL ● NEXT TO YOUR CHOICE.
- Use only the marking device provided or a number 2 pencil.
- If you make a mistake, don't hesitate to ask for a new ballot. If you erase or make other marks, your vote may not count.

| PRESIDENT AND VICE PRESIDENT | STATE | COUNTY |
|---|---|---|
| PRESIDENT AND VICE PRESIDENT<br>(Vote for One) | COMMISSIONER OF GENERAL LAND OFFICE<br>(Vote for One) | DISTRICT ATTORNEY<br>(Vote for One) |
| ○ Gordon Bearce<br>Nathan Maclean REP | ○ Sam Saddler REP | ○ Corey Behnke REP |
| ○ Vernon Stanley Albury<br>Richard Rigby DEM | ○ Elise Ellzey DEM | ○ Jennifer A. Lundeed DEM |
| ○ Janette Froman<br>Chris Aponte LIB | COMMISSIONER OF AGRICULTURE<br>(Vote for One) | COUNTY TREASURER<br>(Vote for One) |
| CONGRESSIONAL | ○ Polly Rylander REP | ○ Dean Caffee REP |
| UNITED STATES SENATOR<br>(Vote for One) | ○ Roberto Aron DEM | ○ Gordon Kallas DEM |
| ○ Cecile Cadieux REP | RAILROAD COMMISSIONER<br>(Vote for One) | SHERIFF<br>(Vote for One) |
| ○ Fern Brzezinski DEM | ○ Jillian Balas REP | ○ Stanley Saari REP |
| ○ Corey Dery IND | ○ Zachary Minick DEM | ○ Jason Valle DEM |

**FIG. 5.** Optical scan bubble example.

these ballots can be seen in Tibbetts and Mullis, 2008), we presented a best case scenario in this study, utilizing ballots that are clearly marked, with no need for interpretation by election auditors.

## PREVIOUS AUDITING AND COUNTING RESEARCH

While assessments of counting procedures have been conducted in numerous states and jurisdictions (Hall, 2008b; Atkeson et al., 2008; Olson, 2009; Halvorson and Wolff, 2007; Alvarez, Katz and Hill, 2005; Bertelsen, 2007; Georgia Secretary of State, 2007), these counting procedures have not been exhaustively studied under controlled conditions with known ballot counts. Laboratory study, in which the true count and condition of the ballots is known objectively, allows us to precisely quantify errors and discrepancies between different styles of counting and different ballots.

In one of the few analyses of hand-counting accuracy, Ansolabehere and Reeves (2004) found that hand-counted ballots generally have higher rates of adjustment upon recount. This analysis, using data from 1946–2002 in New Hampshire, unfortunately uses the hand-counted total in a recount as the truly accurate count, although it may also contain error. Therefore, discrepancies between the original count and the recount could be introduced in either count. Nevertheless, their results suggest that hand-counting can indeed introduce error. More evidence of the fallibility of hand counts can be seen in Atkeson et al. (2008). The authors examined discrepancies between two different machine counts as well as two different hand counts of over 47,000 ballots from the 2006 New Mexico election. While the true count was unknown, the authors found in a real-world setting with two and three-person audit teams that a number of counts result in different aggregate count totals. They found that between 52% and 76% of batches resulted in exact agreement between machine and hand counting methods, depending on whether the ballots were cast on Election Day, early, or absentee. Interestingly, they also found that the two hand counts only result in exact agreement between 45% and 64% of the time. These examinations of real-world recounts provide some evidence that ballot counting is not an error-free process; however, the influence of different types of counting procedures as well as ballot types in controlled settings with known true counts has not previously been tested.

Previously, authors have examined the auditability of thermal VVPAT ballots (Goggin and Byrne, 2007), as well as compared the auditability of optical scan, Voter Verified Video Audit Trail (VVVAT), and VVPAT ballots (Goggin et al., 2008). While the results from both studies are quite similar, Goggin et al. (2008) found that individual auditors only provided correct counts 45.0%, 65.0%, and 23.7% of the time for thermal VVPAT, optical scan, and VVVAT ballots, respectively. With regard to subjective responses of the auditors to the technologies, the studies found participants reported numerous problems with the three ballot types, and suggested several improvements. Furthermore, the auditor's ratings of confidence in the accuracy of the counts they provided was uncorrelated with their actual accuracy. However, one key limitation of both of these studies was the lack of group counting procedures, with only individual auditors used for counting the different styles of ballots. Ideally, counting by audit teams should help reduce the error rates observed in these studies by introducing redundancy and cooperative checking of the counts.

Additionally, this analysis assumes the counting procedure as an independent part of an audit process. By providing a best-case scenario in terms of ballot clarity and interpretation, we can reduce the amount of noise in the data and extract baseline efficiency, or timing data for counting specific numbers of ballots utilizing each method. If problems with sampling, chain-of-custody of the ballots, disputes over specific ballots, or reporting of the counts introduce other errors, we do not capture it here, nor do we capture the additional time and other costs that result from these other parts of the post-election audit procedure. Therefore, if anything, we expect that our data underestimate both the time necessary for audits and error rates, as a more complex, real-world audit would likely introduce more error and take more time to resolve discrepancies and legal challenges from third party observers.

## METHODS

### *Participants*

A total of 108 individuals participated in the current study, with 15 groups of four participants each utilizing the "read-and-mark" procedure, and 16 groups of three participants each utilizing the "sort-and-stack" procedure. The average age was 59 (SD = 17), with 47 male and 55 female participants. Six participants declined to provide demographic information.

Our sample was quite well-educated. One-third (33.3%) possessed some form of post-graduate degree, 36.1% were college graduates, 19.4% had some college experience, while 5.6% possessed only a high school diploma. No participants possessed less than a high school education. Some of our sample also possessed prior election experience. 25.9% had previously worked as a poll worker, working in an average of 6.2 (SD = 8.6) elections. While this is not directly the same as experience as a post-election auditor, it nevertheless implies some familiarity with election procedures.

The participants were recruited from two sources. Members of the public were recruited through newspaper advertisements (n = 53), and were compensated $20 for their participation. Additionally, volunteers from an educational program for older adults (n = 55) participated with no compensation. Participants from the two groups were significantly different in age, t(106) = 7.61, p < .001, with the paid public having an average age of 49.8 (SD = 17), while the volunteers were significantly older with an average age of 69.3 (SD = 8.3). To ensure participants from the two groups did not perform differently, we controlled for effects of the recruitment method and age in our models. We will discuss this more in the design section below.

### *Design*

In the current study, three between-groups independent variables were manipulated, with one additional variable manipulated within-group, as each group counted two separate races. The first between-groups variable was that of counting method. As detailed above, groups of four or three participants were assigned to either use the read-and-mark counting procedure or the sort-and-stack procedure. The second between-groups variable manipulated was that of ballot type. Groups received 120 thermal VVPATs, Prime III VVPATs, or optical scan ballots for the counting task. Finally, the third between-groups variable that was manipulated was the number of rejected, or spoiled ballots contained within each ballot allotment (4 or 8 spoiled per 120 ballots). Additionally, because subjects were recruited from two different pools with significantly different ages, we control for this in all analyses to ensure no bias is introduced. This is done by utilizing an ANCOVA with the mean age for each group included as a continuous covariate. We use mean group age and not recruitment method because it is a more fine-grained measure of the group differences, and should better capture any group differences.

The within-groups variable manipulated was that of closeness of the contest. Because each group counted two contests on the ballot, one contest was lopsided (30% margin of victory), while one contest was relatively close (5% margin of victory). The order in which groups counted the two contests was randomized.

Participants were assigned to a counting group based on scheduling convenience for members. The experimental condition was randomized for each group, however. Similarly, the roles within each group, specifically for the read-and-mark condition, were randomly assigned.

Four quantitative dependent variables were measured in this study, in addition to numerous open-response questions about ballot design, procedure design, and group interaction and the other team members. Three of the quantitative variables measured correspond directly to the three usability metrics we discussed previously: effectiveness, efficiency, and satisfaction. Additionally, we asked participants to rate their confidence in the accuracy of their counts of the ballots on a 5-point Likert scale. Effectiveness has been quantified in terms of error rates, or how the group's counted totals differed from the accurate election results. For efficiency, the groups were timed for each contest counted. Finally, for satisfaction, a subjective metric, we utilized the System Usability Scale (SUS) (Brooke, 1996), which consists of a standard ten-question battery about a system, which in this case was the counting procedure. This commonly used scale has been utilized in previous assessments of satisfaction of voting systems used by voters (e.g., Everett et al., 2008), as well as auditing usability (Goggin and Byrne, 2007; Goggin et al., 2008).

### *Materials*

For the study, the ballots counted were prepared in a controlled environment with all prior vote counts known. For all three types of ballots, 120 fully completed ballots were prepared, each containing 27 contests, comprised of 21 electoral offices and 6 propositions. All material on the ballots was fictional, yet realistic, and was originally prepared by Everett, Byrne, and Greene (2006). Of these 27 contests on each ballot, only two were counted by each team of auditors. Participants counted both the "US Representative District 7" race, appearing third from the top of the ballot, as well as the "County District Attorney," appearing in the 16th position from the top. Additionally, a level of "roll-off," or the increasing rate at which voters under-vote as a function of ballot length, was added to make the ballots appear more realistic. The under-vote rates were 9% and 15%, respectively, for the two races noted above. These rates were based on the findings of Nichols and Strizek (1995). The content of all three types of ballots was identical, with only the formatting being different.

The thermal VVPAT ballots, identical to those used in Goggin and Byrne (2007), as well as Goggin et al. (2008) were prepared to look as similar as possible to real VVPATs from commercial DREs, as well as meet VVSG standards. A complete spool of thermal VVPATs containing all 120 ballots was given to each audit team. While some jurisdictions leave the ballot spool intact through the counting process, we had our participants separate the ballots using a scissors for all conditions because the "sort-and-stack" method requires the ballots be separate. Because "rejected," or invalid ballots, can be wound into the spool of valid ballots if voters reject them while casting their vote, a notation is printed at the bottom of every ballot specifying whether the ballot has been accepted or rejected by the voter. Participants must therefore separate the valid from invalid ballots as they count.

The Prime III VVPATs, printed on legal sized (8.5 inch by 14 inch) paper, were quite similar in layout to the thermal VVPATs, except for the obvious difference in paper size. Like the thermal VVPATs, the Prime III ballots contained only the choice made by the voter, printed underneath the heading of each race. Similarly, both the thermal and Prime III VVPATs placed the phrase "No Vote Cast" underneath the heading of each race if it was skipped or under-voted by the voter.

The optical scan ballots, also on legal sized paper, were identical to those first used by Everett, Byrne, and Greene (2006), and previously used in an auditing context in Goggin et al. (2008). Unlike the two VVPAT ballots, in which all the races were contained in a single vertical column, the optical scan ballot places the 27 contests on the ballot in three columns. In order to make the optical scan ballots similar to the VVPATs in that they contain voided, or rejected, ballots, some ballots were intentionally over-voted to render them invalid.

A different set of instructions was prepared for each of the six permutations of the counting methods and ballot types. These instruction sets used common language, and were made to be as similar in length and depth as possible. Each set of instructions spoke about features of the ballot type the participants would be counting, as well as detailing the specific procedure for counting. The instructions consisted of a single double-spaced page, with the opening paragraph describing the task in general terms, along with key features of the ballots, followed by a bulleted list of steps for each part of the counting process, followed by a few key emphasis points. A second page, including graphics of excerpted parts of the ballots noted key points, including the location of the accepted or rejected notation on the VVPAT ballots, as well as visual examples of an over-voted or under-voted race on an optical scan ballot, with examples provided for only the specific ballot type assigned to the audit group.

### *Procedure*

First, participants were seated around a large table (3 foot by 6 foot, or larger) where the audit would take place. Participants sat wherever they pleased, and were encouraged to use the entire table to help organize and count the ballots. In the case of the read-and-mark procedure, both talliers sat next to each other on the same side of the table, and roles were randomly assigned before the experiment began. Prior to the instructions being given to participants, each participant was told their role for the study (e.g., caller, tallier, witness).

Following the seating and role assignment, a series of verbal instructions were given to the participants by the experimenter. These instructions told the participants of what the study entailed, telling them about the type of ballots they would be counting, as well as the method they would be using. After the short verbal introduction, participants were given the written instructions detailed above. Participants were given ample time to read the written instructions, after which the key points were verbally emphasized. Sample full-size ballots were shown to all participants, with specific locations of necessary notations emphasized. The experimenter encouraged the participants to ask questions both prior to and during the experiment if necessary. The participants were told that the experimenter would act similar to an election observer, watching the process, and was available to provide help with the instructions throughout the counting.

Once the participants indicated they understood the instructions and were ready, a stack or spool of the ballots was given to the group, along with tally sheets for the selected race. The tally sheets specified the contest to be counted, and the contest was verbally specified. Once the first contest was counted, the experimenter collected the tally sheets and provided the group with another tally sheet for the second contest to be counted. Participants counted the two races on the same set of ballots. Once the second contest was counted, tally sheets were collected, and the participants were given a questionnaire, which completed the experiment. This questionnaire contained the SUS, questions about participant confidence in accuracy, open-response questions about numerous aspects of the experiment, as well as demographic information. As the experimenter was unaware of the true ballot counts, at no point in the experiment were the correct ballot counts revealed to the participants.

## RESULTS

In addition to efficiency, accuracy, satisfaction, and confidence, questionnaires prompted participants for written comments about the counting procedure, ballots, and the instructions and experimental materials. We will first discuss the four quantitative metrics, followed by a discussion of the participants' comments from the questionnaires, as well as our descriptive analysis of the process from the experimenter's recorded observations. The efficiency and accuracy data are analyzed at the level of each group of auditors, while the satisfaction, confidence, and open-response data are analyzed at the individual level. One group, utilizing the sort-and-stack method, was excluded from all further analyses due to a failure to follow specified procedural directions; this group declined to count the ballots using the provided instructions, instead relying on methods of their own choosing, even after experimenter intervention.

### *Accuracy (effectiveness)*

Because accuracy is easily the most important metric for any audit system, we present it first. Errors can manifest in several ways: over-counting a candidate's ballots, under-counting their ballots, or attributing ballots to the wrong candidate. Therefore, we evaluate over-counts and under-counts separately, while also analyzing a total error measure that is the absolute value of the difference from the true count. Unfortunately, because we only have aggregate vote totals, we cannot exactly know the percentage of ballots that were incorrectly counted for the opposite candidate (a "wrong-vote" error), as these errors could cancel each other out.

Because there are two candidates per race counted, we can evaluate errors at both the candidate level and the race level. At the candidate level, errors result from differences between a candidate's true vote total and the count for that candidate provided by the group; we term this the candidate total error for each candidate. At the race level, errors represent differences between the true total number of valid ballots (120 in all cases) and the total counts provided by the group, which we term the ballot total error. Because audit groups could get the race level (i.e., total ballot number) count correct but attribute ballots to the incorrect candidate, we need to evaluate both types of error. Both these counts were divided by the total number of ballots (120 in all cases) and multiplied by 100 in order to yield error percentages rather than raw counts.

Overall, 40.0% (SE = 6.4%) of groups provided an incorrect total number of valid ballots, and 46.7% (SE = 6.5%) of groups provided an incorrect count for at least one of the four candidates. The average error percentage for the total number of valid ballots is 1.2% (SE = 0.28%), and the average error percentage for candidate counts is 1.4% (SE = 0.30%). The ballot total error proportions separated by under-counts, over-counts, and the total error by ballot type and counting procedure are shown below in Table 1.

It is worth noting the clear variation in types of errors in ballot total counts across the different types of ballots and counting procedures. Because of the large variation in error proportions, the differences in total error between ballot type are not significant, F(2, 24) = 0.57, p = .58. While there is no significant difference in the total error proportions between counting procedure, F(1, 24) = 0.79, p = .38, the read-and-mark procedure does produce a significantly lower under-count rate than the sort-and-stack procedure, F(1, 24) = 4.50, p = .044.

Turning to candidate total error, we find similar results, as shown above in Table 1. Again, due to the large standard errors, ballot type does not produce significantly different error proportions, F(2, 24) = .95, p = .40. Similarly, the difference in total error proportions between counting procedures is not significant, F(1, 24) = 2.5, p = .127. However,

Table 1. Ballot Total and Candidate Total Error by Ballot Type, Procedure, and Error Type

| | | *Total Error* | | *Under-Count* | | *Over-Count* | |
|---|---|---|---|---|---|---|---|
| | | *Mean* | *SE* | *Mean* | *SE* | *Mean* | *SE* |
| *Ballot Total Error* | | | | | | | |
| Ballot Type | Optical Scan | 0.95% | (.328%) | 0.65% | (.310%) | 0.30% | (.179%) |
| | Thermal VVPAT | 1.80% | (.408%) | 0.40% | (.255%) | 1.40% | (.400%) |
| | Prime III VVPAT | 0.95% | (.671%) | 0.10% | (.100%) | 0.85% | (.670%) |
| Procedure | Read-and-Mark | 0.96% | (.289%) | 0.07% | (.050%) | 0.89% | (.292%) |
| | Sort-and-Stack | 1.47% | (.469%) | 0.66% | (.248%) | 0.81% | (.439%) |
| *Candidate Total Error* | | | | | | | |
| Ballot Type | Optical Scan | 1.87% | (.678%) | 1.03% | (.568%) | 0.85% | (.426%) |
| | Thermal VVPAT | 0.95% | (.218%) | 0.23% | (.131%) | 0.73% | (.196%) |
| | Prime III VVPAT | 1.25% | (.565%) | 0.45% | (.265%) | 0.80% | (.517%) |
| Procedure | Read-and-Mark | 0.48% | (.137%) | 0.04% | (.025%) | 0.45% | (.137%) |
| | Sort-and-Stack | 2.13% | (.538%) | 1.03% | (.393%) | 1.09% | (.414%) |

the difference in under-count proportions between counting procedures is closer to significance, F(1, 24) = 3.43, p = .076.

### *Efficiency*

Efficiency, or timing data, was analyzed for both races that the groups counted. Unsurprisingly, an ANOVA revealed a significant effect of order, F(1, 19) = 26.56, p < .001, with the first race counted taking significantly longer (M = 21.7 minutes, SD = 12.9) than the second race (M = 14.6, SD = 8.9). There was no significant main effect of ballot type; however, the trend present in the data is similar to that found in Goggin et al. (2008), F(2, 19) = 2.26, p = .132, with the thermal VVPAT requiring more time to count than the other two ballot types in both races.

The ballot auditing times were significantly different between the types of counting procedure, F(1, 19) = 17.2, p = .001, with the sort-and-stack method taking significantly longer than the read-and-mark method in both the first and second counts. This effect is clear in both the first and second counts, as shown in Figure 6.

### *Satisfaction*

The System Usability Scale, or SUS, is a 10-question generic battery about a user's satisfaction with a system. These questions, both positively and negatively scored, form a 100-point scale, which is often considered similar to a grading scale. That is, 60 = F, 90+ = A.

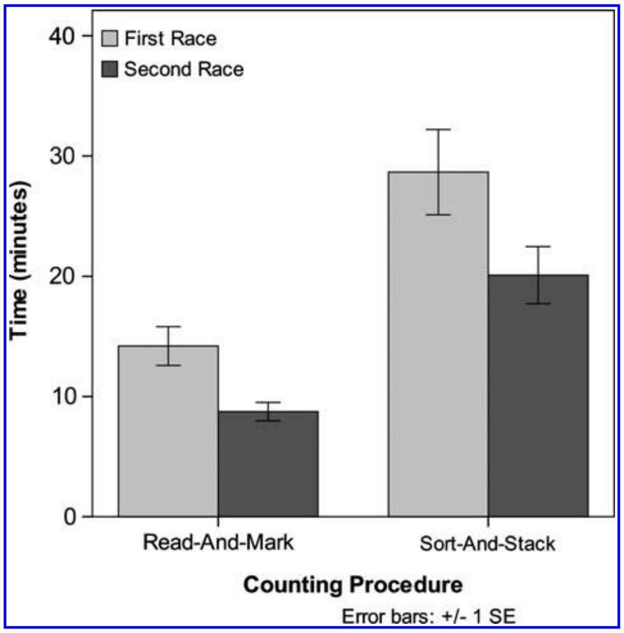


**FIG. 6.** Ballot auditing time by counting procedure and order.

With regards to counting procedure, we find that the read-and-mark procedure received an average SUS score of 79.2 (SD = 16.9), while the sort-and-stack procedure received an average SUS score of 57.6 (SD = 21.2). This difference is significant, F(1, 90) = 39.81, p < .001. The comparison of SUS scores between ballot type is not significant, F(2, 90) = 3.051, p = .052. Finally, a two-way interaction between procedure and ballot type is significant, F(2, 90) = 3.36, p = .039. A closer analysis of this interaction reveals that it results from the optical scan ballot's relatively consistent scores between counting methods, while the thermal VVPAT and the Prime III VVPAT are rated notably worse in the sort-and-stack condition.

### *Confidence*

Similar to above, a factorial ANOVA was conducted to examine the effect of both counting procedure and ballot type on participant's confidence scores on a 1–5 Likert scale, with 5 representing strong confidence in the accuracy of the audit performed. The overall confidence ratings were quite high—the average confidence rating was a 4.49 (SD = .856) across all conditions. In fact, nearly all respondents marked a 4 or a 5, with only a few participants selecting a 1 response indicating low confidence in their accuracy. This is not entirely surprising, as respondents were never told of the true counts at any point in the experiment, just as they would not necessarily know the true count of a stack of ballots in a real election.

We found that the read-and-mark procedure resulted in significantly higher confidence (M = 4.68, SD = .741) than the sort-and-stack procedure (M = 4.24, SD = .933), F(1, 91) = 5.47, p = .021. The effect of ballot type on confidence was not significant, F(2, 91) = .72, p = .49.

### *Participant comments*

The questionnaire at the end of the experiment asked six open-ended questions about different aspects of the experiment, which provided us with valuable comments and suggestions for the procedure and ballot design. Response rates to these six questions ranged from 96.3% to 78.5%.

Some of these questions warranted quantitative analysis, while others merely provided helpful comments about problematic points within the procedures or with the ballots. For instance, when asked whether any discrepancies in counts between group members had to be resolved, 64.6% of participants in the sort-and-stack method noted that they had to resolve at least one discrepancy, while only 26.7% of participants using the read-and-mark method reported having to resolve at least one discrepancy. This difference is statistically significant, $\chi^2(1, N = 99) = 20.07$, $p < .001$. From observation of the experimental protocol, this difference is significant not only in magnitude, but in the amount of effort needed to resolve a discrepancy. With the read-and-mark procedure, resolving a discrepancy often entailed simply rolling back to the last count of ten where the two talliers agreed. With the sort-and-stack procedure, each of the three participants often had to entirely recount the stacks in order to resolve the discrepancy.

When asked about the procedure they used, 90% of participants utilizing the read-and-mark method noted that they no problem implementing it, while only 81.4% of the participants using the sort-and-stack method were able to implement the counting procedure without at least one difficulty. This difference is not statistically significant, $\chi^2(1, N = 103) = 1.58$, $p = .21$. Between the optical scan, VVPAT, and Prime III ballots, 82.4%, 85.8%, and 91.2% of participants noted no problems in implementing the counting procedure. These counts are not significantly different from one another, $\chi^2(2, N = 103) = 1.15$, $p = .56$.

When asked for comments about the different ballot types, participants were quite willing to provide suggestions, with 83.5% of participants responding to the question. Suggestions obviously varied by the type of ballot counted, and many suggestions are procedurally difficult to implement, if not legally impossible. Helpful suggestions, which were given for all three types of ballots included: better marking and separation of races on the ballots, with larger font size and bolded text for the name of each candidate. Thermal VVPAT ballots drew much criticism, with many participants noting the need for a better mechanism by which to separate the ballots from the spool. Some participants suggested perforated or separate sheets of paper in the machine, in addition to using better quality paper that is less slippery and flimsy like the thermal paper used. Several participants commented that rubber thimbles and other clerical tools might be helpful in separating the ballots more easily.

As for procedural suggestions, few had useful comments, as most had no noticeable problems implementing the procedure. However, our error rates suggest that many actually did have problems, even if they closely followed the procedure. Some of the helpful comments included ways to further separate the labor in the read-and-mark method, specifically blinding the talliers to the two individuals counting the ballots so that the separation and handling of the ballots is not distracting, and only auditory attention is required. Several comments were highly critical of the procedures used, writing “don’t ever use this system,” as well as “have a machine do it instead.” From these responses, as well as the quantitative data from the SUS scale, we can see that not all individuals are satisfied with their performance on the ballot-auditing task. Additionally, we find that more problems with the procedure, as well as more discrepancies occurred when participants were utilizing the sort-and-stack method.

### *Experimenter observation*

From our observation of the procedure during the experiments, there are several possible reasons for the observed differences between counting procedures and ballots. A researcher sat next to each counting team during the task and recorded missteps in the procedure and where error could be introduced and inefficient use of time.

Regarding the counting procedure, it appeared that the read-and-mark procedure generated less confusion at the beginning of the procedure due to the clear division of labor and the simplicity of the task given to each participant. Whereas the sort-and-stack procedure had each individual doing multiple tasks over the course of the audit, the read-and-mark procedure tasked each individual with a single simple task to do over the entire course of the count. While there is a danger of the simple tasks becoming overly repetitive, the read-and-mark procedure essentially turned the count into a vigilance task for each individual, while the sort-and-stack procedure required more learning and more higher-order reasoning about implementing the procedure.

With the labor of the task divided among the four participants in the read-and-mark procedure, there were fewer instances of confusion, since each participant only had one task and did not have to coordinate with the other group members. Specifically, with the sort-and-stack procedure, participants could become confused about the different stacks of paper on the table and which had been counted and which stack belonged to which candidate. With the read-and-mark procedure, the process of counting only occurs once with multiple observers, preventing confusion about the ballot's status in the future. Additionally, with the read-and-mark procedure, errors in ballot interpretation could be caught immediately by the witness, whereas the sort-and-stack procedure required an error created by one participant to be sorted out and reconciled later.

With the sort-and-stack procedure, constant communication and cooperation with other group members was necessary to achieve accurate counts and break up the labor. This division of labor, similar to how production facilities have increased quality control through the implementation of "assembly-line" procedures, seems to be beneficial for a group counting task. We believe this division of labor is largely responsible for the differences in nearly all the metrics between the two counting procedures. While experienced ballot auditors would have more time to learn the sort-and-stack procedure before it is implemented, we do not believe this is an adequate solution to this problem. Training of poll workers and election officials is anything but standardized across the country, and removing as much uncertainty and subjective decisions from those implementing the audit is preferable.

From our observation of the handling of the ballots, clear problems were presented with the handling of the thermal VVPAT ballots. While these ballots were not significantly different on several of the metrics, the participants under both procedures had problems separating the ballots and keeping them properly collated. Because the ballots are wound onto spools and stored this way, the ballots have a tendency to curl up and are slippery to handle. This makes counting difficult under both procedures, whether sorting them into stacks or merely reading them aloud and then passing them off, their nature makes the vigilance task of counting more difficult. The subjective comments from participants tell the story here: many of the negative comments about ballot design were directed at the thermal VVPAT ballot. While subtle differences in ballot design, including changing font sizes, labels, and other formatting might produce subtle differences in the ability of the ballots to be easily audited, it appeared the largest difference was made by the paper itself. With the Prime III VVPAT and the optical scan ballot, participants were easily able to handle the legal-sized paper ballot under both procedures.

## DISCUSSION

While we cannot exhaustively evaluate all the election auditing procedures in use around the United States today, this study provides a starting point for objectively evaluating the usability of different methods and their relation with the type of ballots in use. Election auditing is a complex process with many moving parts. Even by focusing on one aspect, the manual audit procedure, as we have done, there are still many variables that can influence the outcome of an audit (see Hall, 2008c for a lengthy discussion). While some previous research has examined election auditing from end-to-end, we think it is essential to break the audit procedure down into its component parts to more precisely analyze how the procedures can fail and introduce error into election counts. As this study has demonstrated, even with the relatively simple task of manually counting ballots, error is ever-present. Furthermore, our findings suggest that the ballot design and counting methods used differentially impact the accuracy of these counts.

While the efficiency data suggests that participants in the study did become quicker at counting over time due to increased familiarity with the procedure, there is no evidence this made a difference in error counts. While increased familiarity with the audit procedure might improve their efficiency in a real election context, we doubt that highly experienced auditors would be able to reduce the observed error rates significantly. As noted above, the counting task is not one that requires experience or high skill, it simply requires a large commitment of concentration and vigilance for extended periods of time. This current study also only required participants to count two races of 27 on the ballots. While we have no reason to suspect results would be different for full ballot recounts, different procedures for this type of recount may be more efficient.

We find similar results to Goggin and Byrne (2007) and Goggin et al. (2008) in terms of the ballot type effects. The thermal VVPAT spools are not well liked by participants, and the additional task of separating the ballots proves quite costly in terms of efficiency, particularly in the sort-and-stack procedure. The introduction of the Prime III VVPAT, which has not been used in previous auditability studies, provides an interesting center point between the usability of thermal VVPAT spools and optical scan ballots. The Prime III VVPAT ballots, because of their normal paper size seem to be handled more easily by participants, but the layout still appears unfamiliar to many participants. The optical scan ballots, however, remain a top performer, likely because of the ubiquity of the formatting and its use in other institutional contexts, such as standardized testing.

A finding common to nearly all the different quantitative and qualitative metrics we use is that well-specified and consistent procedures help improve audits. While errors are not necessarily completely eliminated, nor are audits cost-free, but highly specific procedures for manual auditing help reduce confusion and create a replicable, efficient audit. Furthermore, we find that division of labor to specific tasks of an audit, rather than simple serial counting of ballots to check for accuracy, generally produces more efficient, and less confusing audits. Both of the procedures utilized in this study were very specific in their demands, and ambiguous ballots and other real-world problems were not present. Because of this, our estimates of error and efficiency should represent best-case scenarios.

## CONCLUSION

Overall, this study provides valuable quantitative and qualitative evidence that manual post-election auditing is not an error-free process. Depending on the procedure used, as well as the type of ballot counted, manual audits can vary in their accuracy and efficiency, as well as their appearance of validity to the auditors and outside observers. While many argue manual audits are the "gold standard" by which we must evaluate computerized ballot totals due to the insecure nature of such machines, we must be careful to remember that even the most basic tasks performed by humans can and do introduce error into the process.

## ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under grants #CNS-0524211 (the Accurate Center) and #IIS-0738175. 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

Alvarez, R. M., Katz, J. N., and Hill, S. A. (2005). Machines versus humans: The counting and recounting of pre-scored punchcard ballots. Caltech/MIT Voting Technology Project Working Paper #32.

Ansolabehere, S. and Reeves, A. (2004). Using recounts to measure the accuracy of vote tabulations: Evidence from New Hampshire Elections 1946–2002. Caltech/MIT Voting Technology Project Working Paper #11. http://www.vote.caltech.edu/drupal/files/working_paper/vtp_wp11.pdf.

Atkeson, L. R., Alvarez, R. M., and Hall, T. E. (2008). The New Mexico 2006 post election audit report. http://www.unm.edu/ ∼ atkeson/documents/NM_Audit_Report.pdf.

Bertelsen, J. (2007). 1% Manual tally observer report, Congressional district 11. http://www.countedascast.com/docs/CD11_Manual_Tally_Report_Jan01.pdf.

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

Cross, E.V., Rogers, G., McClendon, J., Mitchell, W., Rouse, K., Gupta, P., Williams, P., Mkpong-Ruffin, I., McMillian, Y., Neely, E., Lane, J., Blunt, H., and Gilbert, J.E. (2007). Prime III: One Machine, One Vote for Everyone. VoComp 2007, Portland, OR, July 16, 2007.

Duffy, J., Turrill, N., Gracely, E., Halvorson, M., Hankins, B., Miller, K., Pierson, L., and Simons, B. (2009). Report on Election Auditing by the Election Audits Task Force. http://www.lwv.org/Content/ContentGroups/Membership/ProjectsTaskforces/Report_ElectionAudits.pdf.

ElectionAudits.org. (2008). Principles and best practices for post-election audits. http://electionaudits.org/principles.html.

Election Data Services. (2008). Nation sees drop in use of electronic voting equipment for 2008 election—a first. October 17th. http://www.electiondataservices.com/images/File/NR_VoteEquip_Nov-2008wAppendix2.pdf.

Everett, S. P., Byrne, M. D., and 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., and Torous, T. (2008). Electronic voting machines versus traditional methods: Improved preference, similar performance. Human Factors in Computing Systems: Proceedings of CHI 2008 (883–892). New York: ACM.

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.

Goggin, S. N. and Byrne, M. D. (2007). An Examination of the Auditability of Voter Verified Paper Audit Trail (VVPAT) Ballots. Proceedings of the 2007 USENIX/ACCURATE Electronic Voting Technology Workshop. Boston, MA.

Goggin, S. N., Byrne, M. D., Gilbert, J. E., Rogers, G., and McClendon, J. (2008). Comparing the auditiability of optical scan, voter verified paper audit trail (VVPAT) and video (VVVAT) ballot systems. Proceedings of the 2008 USENIX/ACCURATE Electronic Voting Technology Workshop. San Jose, CA.

Hall, J. L. (2008a). Procedures for California's 1% manual tally. University of California at Berkeley, School of Information. http://josephhall.org/procedures/ca_tally_procedures-2008.pdf.

Hall, J. L. (2008b). Improving the security, transparency and efficiency of California's 1% manual tallyprocedures. Proceedings of the USENIX/ACCURATE Electronic Voting Technology Workshop. San Jose, CA.

Hall, J. L. (2008c). Policy mechanisms for increasing transparency in electronic voting. Unpublished Doctoral dissertation. University of California, Berkeley. http://josephhall.org/papers/jhall-phd.pdf.

Halvorson, M. and Wolff, L. (2007). Report and analysis of the 2006 post-election audit of Minnesota's voting systems. Citizens for Election Integrity. http://www.ceimn.org/files/CEIMNAuditReport2006.pdf.

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

McMillian, Y., Williams, P., Cross, E.V., Mkpong-Ruffin, I., Nobles, K., Gupta, P., and Gilbert, J.E. (2007). Prime III: Where Usable Security and Electronic Voting Meet. Usable Security (USEC '07), Lowlands, Scarborough, Trinidad/Tobago, February 15–16, 2007.

New Hampshire Department of State. (2006). New Hampshire Election Procedure Manual: 2006–2007. http://www.sos.nh.gov.

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

Ohio Joint Audit Working Group. (2008). Consideration of Ohio audits in 2008. http://www.caseohio.org/Documents/Reports/Ohio_Audit_White_Paper_Feb_2008.pdf.

Olson, R. (2009). Ohio 2008 post election audit review. CASE Ohio. http://www.caseohio.org/PageDetails/Audits/OH_2008_Audit_Review.pdf.

Norden, L., Burstein, A., Hall, J. L., and Chen, M. (2007). Post-Election Audits: Restoring Trust in Elections. Brennan Center for Justice at The New York University School of Law and the Samuelson Law, Technology and Public Policy Clinic at the University of California, Berkeley School of Law (Boalt Hall). http://www.brennancenter.org/dynamic/subpages/download_file_50227.pdf.

Stevens, A. (2007). Hand counting paper ballots. Presentation at the Democracy Fest Annual National Convention, June 5th. http://www.democracyfornewhampshire.com/files/Hand_count_training_D-fest_July_5_2007.pdf.

Tibbetts, T. and Mullis, S. (2008, December 3). Challenged ballots: You be the judge. Minnesota Public Radio. http://minnesota.publicradio.org/features/2008/11/19_challenged_ballots/.

Tobi, N. (2010). *Hands-on elections*. Wilson, NH: Healing Mountain Publications.

Verified Voting (accessed 10/14/2009). Mandatory Manual Audits of Voter-Verified Paper Records. http://www.verifiedvoting.org.

Voluntary Voting System Guidelines. (2007). Final TGDC Recommendations to the EAC, August 31st. http://www.eac.gov/files/vvsg/Final-TGDC-VVSG-08312007.pdf.

Watson, W., Michaelsen, L. K., and Sharp, W. (1991). Member competence, group interaction, and group decision making: A longitudinal study. Journal of Applied Psychology. 76(6) 803–809.

Webb, N. M. (1982). Group composition, group interaction, and achievement in cooperative small groups. Journal of Educational Psychology. 74(4) 475–484.

Address correspondence to:
*Stephen N. Goggin*
*Department of Political Science*
*University of California, Berkeley*
*210 Barrows Hall #1950*
*Berkeley, CA 94720-1950*

*E-mail:* goggin@berkeley.edu