Can Voters Detect Malicious Manipulation of Ballot Marking Devices?

Matthew Bernhard, Allison McDonald, Henry Meng, Jensen Hwa, Nakul Bajaj*, Kevin Chang, J. Alex Halderman

University of Michigan   *The Harker School

Abstract—Ballot marking devices (BMDs) allow voters to select candidates on a computer kiosk, which prints a paper ballot that the voter can review before inserting it into a scanner to be tabulated. Unlike paperless voting machines, BMDs provide voters an opportunity to verify an auditable physical record of their choices, and a growing number of U.S. jurisdictions are adopting them for all voters. However, the security of BMDs depends on how reliably voters notice and correct any adversarially induced errors on their printed ballots. In order to measure voters' error detection abilities, we conducted a large study (N = 241) in a realistic polling place setting using real voting machines that we modified to introduce an error into each printout. Without intervention, only 40% of participants reviewed their printed ballots at all, and only 6.6% told a poll worker something was wrong. We also find that carefully designed interventions can improve verification performance. Verbally instructing voters to review the printouts and providing a written slate of candidates for whom to vote both significantly increased review and reporting rates—although the improvements may not be large enough to provide strong security in close elections, especially when BMDs are used by all voters. Based on these findings, we make several evidence-based recommendations to help better defend BMD-based elections.

I. INTRODUCTION

The threat of election hacking by hostile nations has prompted a major push to ensure that all voting systems in the United States have voter-verifiable paper trails, a defense recommended by the National Academies [36], the Senate Select Committee on Intelligence [53], and nearly all election security experts. Guided by past research [8], some states and localities are implementing paper trails by deploying ballot-marking devices (BMDs). In these systems, the voter makes selections on a computer kiosk, which prints a paper ballot that the voter can review before inserting it into a computer scanner to be counted [56]. BMDs have long been used as assistive devices for voters with disabilities, and a growing number of jurisdictions are purchasing them for use by all voters [24], [25], [37].

BMDs have the potential to provide better security than direct-recording electronic voting machines (DREs), which maintain the primary record of the voter's selections in a computer database and often lack a voter-verifiable paper trail. Numerous studies have demonstrated vulnerabilities in DREs that could be exploited to change election results (e.g., [11], [23], [31], [35]). In contrast, BMDs produce a physical record of every vote that can, in principle, be verified by the voter and manually audited by officials to confirm or correct the initial electronic results.

However, BMDs do not eliminate the risk of vote-stealing attacks. Malware could infect the ballot scanners and change the electronic tallies—although this could be detected by rigorously auditing the paper ballots [50]—or it could infect the BMDs themselves and alter what gets printed on the ballots. This latter variety of cheating cannot be detected by a post-election audit, since the paper trail itself would be wrong, and it cannot be ruled out by pre-election or parallel testing [51]. Instead, BMD security relies on voters themselves detecting such an attack. This type of human-in-the-loop security is necessary in many systems where detection and prevention of security hazards cannot be automated [18]. However, as several commentators have recently pointed out [7], [20], [51], its effectiveness in the context of BMDs has not been established.

Whether such a misprinting attack would succeed without detection is highly sensitive to how well voters verify their printed ballots. Every voter who notices that their ballot is misprinted and asks to correct it *both* adds to the evidence that there is a problem *and* requires the attacker to change an additional ballot in order to overcome the margin of victory. Consider a contest with a 1% margin in which each polling place has 1000 voters. If voters correct 20% of misprinted ballots, minimal outcome-changing fraud will result in an average of 1.25 voter complaints per polling place—likely too few to raise alarms. If, instead, voters correct 80% of misprinted ballots, polling places will see an average of 20 complaints, potentially prompting an investigation. (We model these effects in Section V.) Despite this sensitivity, voters' BMD verification performance has never before been experimentally measured.

In this paper, we study whether voters can play a role in BMD security. We first seek to establish, in a realistic polling place environment, the rates at which voters attempt to verify their printed ballots and successfully detect and report malicious changes. To measure these, we used real touch-screen voting machines that we modified to operate as malicious BMDs. We recruited 241 participants in Ann Arbor, Michigan, and had them vote in a realistic mock polling place using the ballot from the city's recent midterm election. On every ballot that our BMDs printed, one race was changed so the printout did not reflect the selection made by the participant.

We found that, absent interventions, only 40% of participants reviewed their printed ballots at all, only 6.6% reported the error to a poll worker, and only 7.8% correctly identified it on an exit survey. These results accord with prior studies that found poor

voter performance in other election security contexts, such as DRE review screens [1], [15] and voter-verifiable paper audit trails (VVPATs) [48]. The low rate of error detection indicates that misprinting attacks on BMDs pose a serious risk.

The risks notwithstanding, BMDs do offer practical advantages compared to hand-marked paper ballots. They allow voters of all abilities to vote in the same manner, provide a more user-friendly interface for voting, and more easily support complex elections like those conducted in multiple languages or with methods such as ranked choice [44]. BMDs also simplify election administration in places that use vote centers [56], which have been shown to reduce election costs and lower provisional voting rates [28], [42], as well as in jurisdictions that employ early voting, which can improve access to the ballot [30].

Given these advantages and the fact that BMDs are already in use, the second goal of our study was to determine whether it might be possible to boost verification performance through procedural changes. We tested a wide range of interventions, such as poll worker direction, instructional signage, and usage of a written slate of choices by each voter.

The rate of error detection varied widely with the type of intervention we applied, ranging from 6.7% to 86% in different experiments. Several interventions boosted review rates and discrepancy reporting. Verbally encouraging participants to review their printed ballot after voting boosted the detection rate to 14% on average. Using post-voting verbal instructions while encouraging participants to vote a provided list of candidates raised the rate at which voters reported problems to 73% for voters who did not deviate from the provided slate.

These findings suggest that well designed procedures can have a sizable impact on the real-world effectiveness of voter verification. We make several recommendations that election officials who already oversee voting on BMDs can employ immediately, including asking voters if they have reviewed their ballots before submission, promoting the use of slates during the voting process, informing voters that if they find an error in the printout they can correct it, and tracking the rate of reported errors. Our recommendations echo similar findings about the most effective ways to alert users to other security hazards (i.e., in context [12] and with active alerts [21]) and redirect them to take action.

Although our findings may be encouraging, we strongly caution that much additional research is necessary before it can be concluded that any combination of procedures actually achieves high verification performance in real elections. Until BMDs are shown to be effectively verifiable during real-world use, the safest course for security is to prefer hand-marked paper ballots.

Road Map   Section II provides more background about human factors and security and about previous work studying the role of voter verification in election security. Section III describes our experimental setup, voting equipment, and study design. Section IV presents our results and analyzes their significance. Section V provides a quantitative model for BMD verification security. Section VI discusses the results, avenues for future work, and recommendations for improving the verifiability of BMDs. We conclude in Section VII.

II. Background and Related Work

A. Human-Dependent Security

Elections fundamentally depend on having humans in the loop—as Stark [51] notes, the voter is the *only one* who knows whether the ballot represents their intended vote—and the success or failure of election security has the potential to have history-altering effects. The type of risk posited by Stark, wherein voters do not check their paper ballots to ensure the BMD has correctly represented their selections, is a post-completion error [14], in which a user makes a mistake (or fails to verify the correctness of something) *after* they have completed the main goal of their task. Voters who forget or do not know to verify the correctness of a paper ballot after they have entered their selections on a BMD miss a critical step in ensuring the accuracy of their vote. We therefore explore how to communicate this risk to voters.

Cranor [18] describes five ways that designers can communicate risk to a user who needs to make security decisions:

1) *Warnings*: indication the user should take immediate action
2) *Notices*: information to allow the user to make a decision
3) *Status indicators*: indication of the status of the system
4) *Training*: informing users about risks and mitigations before interaction
5) *Policies*: rules with which users are expected to comply

Implementing indicators that reveal meaningful information to voters about the security status of a BMD would be next to impossible, as security issues are often unknown or unforeseen to the operators. Although voter education about the importance of verification might be an effective form of training, significant coordination would be necessary to enact such a scheme at scale. Therefore, we focus in this study on the effectiveness of warnings issued through poll worker scripts and polling place signage.

A warning serves two purposes: to alert users to a hazard, and to change their behavior to account for the hazard [62]. There are many barriers to humans correctly and completely heeding security warnings. Wogalter proposes the Communication-Human Information Processing (C-HIP) Model [61] to systematically identify the process an individual must go through for a warning to be effective. The warning must capture and maintain attention, which may be difficult for voters who are attempting to navigate the voting process as quickly as possible. Warnings must also be comprehensible, communicate the risks and consequences, be consistent with the individual's beliefs and attitudes toward the risk, and motivate the individual to change—all of which are substantial impediments in an environment with little to no user training and such a broad user base as voting.

To maximize effectiveness, warnings should be contextual, containing as little information as necessary to convey the risk and direct individuals to correct behavior [12], [61]. Voters are essentially election security novices; Bravo-Lillo et al. [12] found that, in the context of computer security, advanced and novice users respond to warnings differently. Most significantly, novice users assessed the hazard *after* taking action, whereas

advanced users assessed the hazard *before* engaging in the activity.

There may be effective ways to improve voter verification performance. Many studies have applied lessons from Cranor, Wogalter, and Bravo-Lillo et al. to help humans make secure choices in different contexts, including phishing [21], [41], browser warnings [2], [46], [52], app permissions [3], [40], and operating system interfaces [13]. In the context of phishing warnings, for example, Egelman et al. [21] found that users were far more likely to heed an active warning, or a warning that disrupted their workflow, than a passive warning. This suggests that similar interventions applied in a polling place may have a significant effect on voters' ability to review and verify their BMD ballots.

Our study contributes to this literature by exploring the effects of several modalities of warnings (oral and visual) on human detection of malicious ballot modification.

B. Voter-Verifiable Paper and Ballot-Marking Devices

A guiding principle in election security is that voting systems should be *software independent* [47]: that is, any software errors or attacks that change the reported election outcome should be detectable. Bernhard et al. [9] note that elections backed by a voter-verifiable paper record are currently the only known way to provide robust software independence. Like BMDs, voter-verifiable paper audit trails (VVPATs) and hand-marked paper ballots are widely used in an attempt to achieve software independence. However, each poses a different set of usability and accessibility challenges.

Hand-marked paper ballots record the voter's selections without the risk of having a potentially compromised computer mediating the process. However, voters often make mistakes when filling out ballots by hand that can lead to them being counted incorrectly or ruled invalid [27]. Moreover, many voters have difficulty marking a paper ballot by hand due to a disability or a language barrier. Ballots in the U.S. are among the most complex in the world, further magnifying these difficulties [38].

VVPAT technology also suffers from noted usability, privacy, and auditability problems [26]. Most implementations consist of clunky printer attachments for DREs that are difficult for voters to read, record votes in the order in which they are cast, and use a fragile paper tape. In laboratory studies, Selker et al. [48] and de Jong et al. [19] found that voters frequently did not review the VVPAT, with Selker finding that only 17% of voters detected changes between the selections they made on the DRE and those printed on the VVPAT. While there has been some criticism of Selker's findings and methodology [45], [49], their results broadly comport with work by Campbell et al. [15] and Acemyan et al. [1] about voters' ability to detect errors introduced in DRE review screens. The latter found that only 12–40% of participants successfully detected such errors.

In part due to the concerns raised by these studies, BMDs have become a popular choice for new voting system deployments in the United States. South Carolina and Georgia, together comprising nearly 9 million voters, recently adopted BMDs statewide [24], [25], as have several counties and cities, including Los Angeles County, the largest single election jurisdiction in the U.S. [58].

There has been vigorous debate among election security experts as to whether BMDs can provide software-independence (e.g., [7], [20], [51], [60]). However, the discussion has yet to be informed by rigorous experimental data. Our work seeks to fill that gap by contributing the first human-subjects study to directly measure the verification performance of voters using BMDs under realistic conditions and with a variety of potential procedural interventions.

III. MATERIALS AND METHODS

Our goals in this work were to empirically assess how well voters verify BMD ballots and whether there are steps election officials can take that will enhance verification performance. To these ends, we conducted a between-subjects study where we tested several hypotheses in a simulated polling place, following the best practices recommended by Olembo et al. [39] for election human-factors research. The study design was approved by our IRB.

We sought to answer several questions, all of which concern the rate at which voters are able to detect that a BMD-printed ballot shows different selections than those the voter picked:

- What is the base rate of error detection?
- Is error detection impacted by:
  - Ballot style?
  - Manipulation strategy?
  - The manipulated race's position on the ballot?
  - Signage instructing voters to review their ballots?
  - Poll worker instructions?
  - Providing a slate of candidates for whom to vote?

In order to answer these questions in an ecologically valid way, we attempted to create an environment that closely resembled a real polling place. Nevertheless, it is impossible for any experiment to fully recreate what is at stake for voters in a real election, and so study participants may have behaved differently than voters do in live election settings. We went to extensive lengths to mitigate this limitation, and we find some data to support that we did so successfully (see Section VI-A). We used real (though modified) voting machines, printers and paper stock from deployed BMD systems, a ballot from a real election, and ballot styles from two models of BMDs. We conducted the study at two city library locations, one of which is used as a polling place during real elections.

A. The Polling Place

To provide a realistic voting experience, we structured our simulated polling place like a typical BMD-based poll site. Three investigators served as poll workers, following the script in Appendix A. Library patrons who were interested in voting began at a check-in table, where they were greeted by Poll Worker A and asked to sign an IRB-approved consent form. Participants were told they would be taking part in "a study about the usability of a new type of voting machine" and instructed

Fig. 1: *Polling Place Setup.* We established mock polling places at two public libraries in Ann Arbor, Michigan, with three BMDs (*left*) and an optical scanner and ballot box (*right*). Library visitors were invited to participate in a study about a new kind of election technology. The BMDs were DRE voting machines that we modified to function as malicious ballot marking devices.

on how to use the equipment, but they were not alerted that the study concerned security or that the BMDs might malfunction.

Each participant received a voter access card with which to activate a BMD and was free to choose any unoccupied machine. There were three identical BMDs, as shown in Figure 1. On the last day of the study, one machine's memory became corrupted, and it was removed from service; all votes that day were recorded on the other two machines.

The BMDs displayed contests in a fixed order, and voters made selections using a touch screen interface. After the last contest, the machines showed a review screen that accurately summarized the voter's selections and highlighted any undervotes. The voter could return to any contest to change the selections. A "Print Ballot" button ended the voting session and caused a printer under the machine to output the paper ballot.

Participants carried their ballot across the polling place to the ballot scanner station, where they inserted them into an optical scanner that deposited them into a ballot box. Poll Worker B was stationed by the scanner and offered instructions if necessary. Next, the poll worker collected the voter access card and asked each participant to complete an exit survey using a laptop next to the scanning station. The survey was anonymous, but responses were keyed so that we could associate them with the voter's on-screen selections, their printed ballot, and poll worker notes.

Poll Worker C, positioned separately from the other stations, acted as an observer. They verified that participants moved through the polling place stations sequentially, noted whether they spent time reviewing their printed ballots, and recorded whether they appeared to notice any abnormalities. The observer was also tasked with noting participant behavior, specifically how the participants completed each step in the voting process and any comments they made. The observer was available to answer participant questions and was frequently the poll worker participants approached upon noticing a discrepancy.

Like in a real polling place, multiple participants could progress through the voting process simultaneously. Occasion-ally a one- or two-person line formed as participants waited to use the BMDs or the ballot scanner.

B. The Voting Machines

BMD voting systems are currently produced by several voting machine manufacturers, the largest of which is ES&S. Over a six month period, we repeatedly attempted to engage ES&S in discussions about acquiring samples of their equipment for this study. However, these attempts were ultimately not fruitful.

Instead, we utilized AccuVote TSX DRE voting machines, which we purchased on eBay and modified to function as BMDs. The TSX was first produced by Diebold in 2003 and is still widely deployed today. At least 15 states plan to use it in at least some jurisdictions in November 2020 [57].

The TSX runs Windows CE and is designed to function as a paperless DRE or a VVPAT system. We developed software modifications that allow it to print ballots in multiple styles using an external printer. This effectively converts the TSX into a BMD — and one we could easily cause to be dishonest — while preserving the original touch-screen interface used by voters.

In order to modify the machine, we built on techniques used by Feldman et al. [23]. We began by patching the firmware so that, when the machine boots, it attempts to execute a program provided on an external memory card. We used this functionality to launch a remote access tool we created, which allowed us to connect to the TSX over a network and perform file system operations, run applications, and invoke a debugger.

The TSXes in our polling place were connected to an Ethernet switch using PCMCIA network adapters. A Python program, running on a computer on the same network, used the remote access tool's API to poll each machine for newly voted ballots. Whenever a ballot was cast, the program parsed the selections, generated a PDF file based on them, and sent it to a printer located underneath the appropriate voting machine. The program could be configured to apply different ballot styles and cheating strategies, depending on the experiment.

For every ballot, the program randomly selected one race to manipulate. In most experiments, selections could be changed in three ways: deselection in a voted-for race, selection in an unvoted-for race, or changing a selection to a different candidate. We ensured that some alteration would take place on every ballot. For example, in a vote-for-one race where the voter had made a selection, the algorithm would choose uniformly from the set of unselected choices plus no selection. One experiment used a different strategy, in which choices could only be deselected.

Both the voter's original selections and the manipulated ballot were logged for later analysis. Each voting session was associated with a unique tracking number, which was printed on the ballot along with a timestamp and encoded as a barcode.

As the final step in the voting process, participants fed their printed ballots into an AccuVote OS optical scanner, a device used to tabulate votes in parts of 20 states [57]. The scanner was intended to add realism to the experiment, but AccuVote OSes are not capable of actually tabulating the ballot styles we used. Therefore, we modified the scanner so that it simply fed each ballot into the ballot box without counting it.

We mounted a barcode reader in a 3-D printed case above the scanner's input tray and positioned it so that it would detect the ballot's tracking barcode. (This setup can be seen in Figure 3.) When the barcode was read, a Raspberry Pi would activate the AccuVote OS's feed motor to pull the ballot into the ballot box. The Raspberry Pi also displayed the ballot tracking number so that poll workers could associate the ballot with the participant's exit survey response and the observer's notes.

C. The Ballot

In order to ensure a realistic voting experience and increase participants' psychological investment in the outcome of the mock election, we used races and candidates from the city's actual ballot for the recent 2018 midterm election. For simplicity, we reduced the ballot to the first 13 races so that ballots would not require duplex printing or multiple pages.

We tested two ballot styles, which are illustrated in Figure 2. One is a regular ballot that shows the entire set of candidates in every race. The other is a summary ballot, which shows only the voter's selections or "NO SELECTION" if a choice is left blank. Most BMDs print ballots that resemble these styles.

The specific visual designs we used mimic ballots produced by two models of BMDs manufactured by Hart InterCivic, which also makes the voting equipment used in Ann Arbor. The regular style is also the same design as the hand-marked paper ballots most Ann Arbor voters use, ensuring that many participants found it familiar. These designs are used in jurisdictions that collectively have over 10 million registered voters [57].

The model of laser printer we used, Brother HL-2340, is certified for use with Clear Ballot's ClearAccess BMD system [43], so we chose paper stock that meets the specifications for ClearAccess [16]. Summary ballots were printed on regular weight 8.5×11 inch letter paper, while regular ballots were printed on Vellum Bristol stock 67 pound 8.5×14 inch paper.
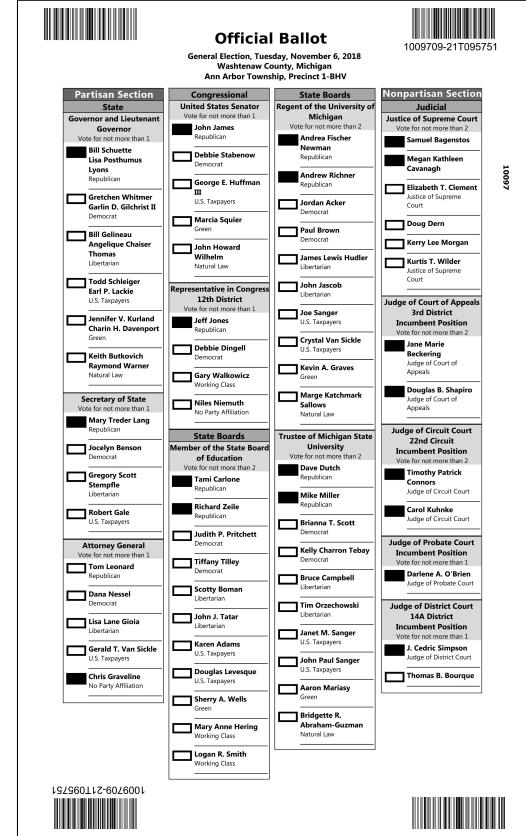
(a) Regular Ballot
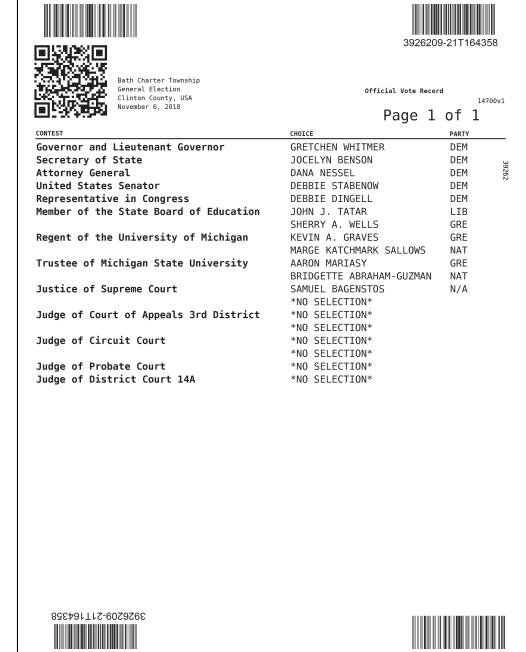
(b) Summary Ballot

Fig. 2: *Ballot Styles.* We tested two ballot styles: (*a*) a regular style, resembling a hand-marked ballot; and (*b*) a summary style, listing only the selected candidates. Both had 13 races from the city's recent midterm election. In one race, determined randomly, the printed selection differed from the voter's choice.

D. Participants and Recruitment

To gather subjects for our study, we approached staff at the Ann Arbor District Library (AADL), who offered space for us to set up our mock precinct. We conducted a total of three days of data collection in July and September 2019 at two library locations: the Downtown and Westgate branches. The Downtown branch, where our study was held for two of the three days, is an official polling location during real elections.

The AADL advertised our study through its social media feeds and offered incentives to patrons for their participation, such as points for a scavenger hunt competition [5] and souvenir flashlights [6]. We also set up a fourth voting machine outside of the mock precinct where kids could vote in an election for mayor of the library's fish tank.[1] Results from that machine were not used as part of this study, but it served as a recruitment tool for parents visiting the library with their children. In addition, we verbally recruited patrons who happened to be at the libraries during our study, using the script in Appendix B.

Participants were required to be at least 18 years of age and to sign an IRB-approved consent form. All data collected, including survey responses and behavioral observations, was completely anonymous. We informed participants that they were not required to vote their political preferences.

E. Experiments

To explore what factors affect voter verification performance, we devised nine experiments to run between subjects. In all experiments, for every participant, one selection that the participant made on the BMD was not accurately reflected on the printed ballot. Every participant within an experiment received the same instructions from the poll workers, following the script and variants in Appendix A.

The first three experiments were designed to measure verification in the absence of protective interventions. They varied the ballot style and manipulation strategy:

E1: Regular ballots   We used the regular ballot style and the default manipulation strategy, in which a selection could be switched, deselected, or selected if left blank by the voter.

E2: Summary ballots   We used the summary ballot style and the default manipulation strategy. As discussed in Section IV, we found no significant difference in error detection between regular ballots and summary ballots, so all subsequent experiments used summary ballots.

E3: Deselection only   To assess the sensitivity of voters to the way their ballots were changed, we limited the manipulation to deselecting one of the voter's choices at random.

Four further experiments tested interventions to determine if they improved error detection. We tried posting a sign and having poll workers give different instructions at various times:

E4: Signage   A sign was placed above the scanner that instructed voters to check their printed ballots, as shown in

[1]Mighty Trisha unexpectedly beat Creepy Bob, leading some Bob supporters to complain that the results were fishy [4].

Fig. 3: *Warning Signage.* One of the interventions we tested was placing a sign above the scanner that instructed voters to verify their ballots. Signage was not an effective intervention.

Figure 3. We designed the sign following guidelines from the U.S. Election Assistance Commission [55].

E5: Script variant 1   During voter check in, the poll worker added this instruction: "Please remember to check your ballot carefully before depositing it into the scanner."

E6: Script variant 2   When the voter approached the scanner, the poll worker said: "Please keep in mind that the paper ballot is the official record of your vote."

E7: Script variant 3   When the voter approached the scanner, the poll worker said: "Have you carefully reviewed each selection on your printed ballot?"

The final two experiments assessed whether reminding participants of their selections during verification improved their performance. We gave voters a slate of candidates for whom to vote that they could carry with them throughout the voting experience. While we refer to this as a slate, a sample ballot that the voter filled in before voting could serve the same purpose. Every voter received the same slate (Appendix C), which was randomly generated and contained an even mix of parties.

E8: Slate with script variant 2   Voters were given the slate. Poll workers encouraged verification with script variant 2.

E9: Slate with script variant 3   Voters were given the slate. Poll workers encouraged verification with script variant 3.

Experiment	N	Were observed examining ballot	Reported error on exit survey	Reported error to poll worker
Without interventions:				
E1: Regular ballots	31	41.9%	6.5%	6.5%
E2: Summary ballots	31	32.3%	6.5%	6.5%
E3: Deselection only	29	44.8%	10.3%	6.9%
Subtotal/Mean	91	39.7%	7.8%	6.6%
With interventions:				
E4: Signage	30	13.3%	3.3%	6.7%
E5: Script variant 1	30	46.7%	13.3%	6.7%
E6: Script variant 2	25	92.0%	16.0%	16.0%
E7: Script variant 3	31	38.7%	19.4%	12.9%
E8: Slate with script variant 2	13	100.0%	38.5%	38.5%
E9: Slate with script variant 3	21	95.2%	71.4%	85.7%
Subtotal/Mean	150	64.3%	24.0%	27.8%

TABLE I: *Verification Performance for Each Experiment.* Without interventions, participants' verification performance was remarkably poor: only 7.8% noted on an exit survey that their ballots had been altered, and only 6.6% informed a poll worker (averaged across experiments). The various interventions we tested had widely different effects, ranging from no significant improvement (**E4**, **E5**) to a large increase in verification success (**E8**, **E9**).

IV. RESULTS

A. Participant Demographics

We recruited 241 participants. The vast majority (220, 91%) indicated that they were native English speakers; 19 reported speaking twelve other native languages, including Hungarian, Korean, and Arabic; and two subjects gave no response. Participants who disclosed their age ranged from 18 to 84 years old, with a mean of 43.7 and a median of 42; 15 subjects did not answer the question. The percentages that follow are out of the total number of responses to each question: Respondents identified as male (84, 35%), female (152, 64%), or other (3, 1%); two did not respond. Subjects reported their ethnicity as Caucasian (187, 80%), Asian (17, 7%), African American (6, 3%), Mexican American/Chicano (5, 2%), and Other Hispanic/Latino (9, 4%); others reported not having any of these ethnic backgrounds (2, 1%) or were multiracial (9, 4%). Participants reported their level of educational attainment as some high school (1, 0.4%), a high school diploma (4, 2%), some college (20, 8%), a two-year degree (10, 4%), a four-year degree (80, 33%), a master's or professional degree (92, 38%), or a doctorate (34, 14%).

Most subjects indicated that they were registered to vote in the U.S. (220, 92%), had voted in a previous election (216, 91%), and had voted in the November 2018 midterm election (209, 87%). However, we note that, historically, 38–45% of non-voters have been found to falsely report having voted [10].

Compared to the population of Ann Arbor at the time of the 2010 census, our participant pool overrepresented Caucasians ($\Delta = 7.6\%$) and underrepresented African Americans ($\Delta = -4.4\%$) and Asians ($\Delta = -8.7\%$) [54]. The study population also overrepresented females ($\Delta = 13\%$) and underrepresented males ($\Delta = -16\%$) [59]. In other reported aspects, participants' demographics resembled the population of Ann Arbor voters (the city is among the most highly educated in the U.S.) [33].

B. Verification Performance

To quantify verification performance, we collected three data points for each participant, which are summarized in Table I. First, an observer noted whether the subject appeared to examine the printed ballot for at least two seconds. Second, the exit survey asked, "Did you notice anything odd about your ballot?", and we recorded whether the subject's response corroborated the discrepancy (i.e., correctly articulated which race was changed). Third, we recorded whether subjects reported the ballot modification to a poll worker. Most experiments saw more participants identify discrepancies in the survey than were reported to poll workers, but these differences were not statistically significant. Where applicable, we refer to participants who by some means reported detecting the discrepancies as "noticers" and those who did not as "non-noticers".

*1) Performance without interventions (**E1–E3**):* With no interventions, we found verification performance to be consistently poor. The three experiments involved 91 participants, and, averaged across the experiments, only 40% of participants examined their ballots, only 7.8% noted the error on the exit survey, and only 6.6% reported it to a poll worker. We did not find significant differences in performance between regular and summary ballots or between the tested attack strategies.

*2) Effectiveness of interventions (**E4–E9**):* The tested interventions resulted in a wide range of effect sizes. Neither signage (**E4**) nor poll worker instructions issued before the participant began voting (**E5**) yielded a statistically significant improvement to any aspect of verification performance. In

contrast, poll worker instructions issued *after* the ballot was printed (**E6** and **E7**) did have a positive effect, boosting reporting rates to 20% on the exit survey and 14% to poll workers (averaged across the experiments).

The largest performance gains occurred when participants were directed to vote using a slate of candidates (**E8** and **E9**). However, only **E9** produced a statistically significant difference in reporting rates (Fisher's exact $p < 0.001$).[2] Averaged across both experiments, reporting rates increased to 55% on the exit survey and 62% to poll workers. **E8**, in which participants were directed how to vote using a slate of candidates, saw detection and reporting rates of 39%, which is similar to results for DRE review screen performance found by Campbell et al. [15] and Acemyan et al. [1], in studies that similarly directed participants how to vote. With script variant 3, the use of a slate produced a significant difference (comparing **E7** and **E9**, Fisher's exact $p < 0.02$) for both review and report, but it did not produce a significant difference using script variant 2 (comparing **E6** and **E8**). This indicates that voters may be sensitive to the specific instructions they receive about reviewing their ballots.

C. Correlates

1) Reviewing the ballot: Reviewing the ballot at all was significantly correlated with error reporting (two-sample permutation test $p < 0.001$ with 10k repetitions). Some interventions do seem to promote reviewing: **E6**, **E8**, and **E9** saw significant increases (Fisher's exact $p < 0.004$), although **E7** did not.

2) Time to ballot submission: Careful verification takes time, so one might expect that participants who noticed discrepancies took more time to cast their ballots. As an upper bound on how long subjects spent verifying, we calculated the time from ballot printing to ballot submission. (Due to clock drift on one of our machines, data from the third day of experiments was unusable, and consequently **E4** and **E7** are excluded from our timing analysis.) As expected, we find that noticers took an average of 121 s between printing and ballot submission (median 114 s), compared to only 43 s for non-noticers (median 32 s). This difference is statistically significant (two-sample permutation test $p < 0.004$, 10k iterations).

We compared the submission times for two sets of experiments: ones with extra instructions to the voter (**E5**, **E6**, **E8**, and **E9**; $N = 84$) and ones without (**E1**, **E2**, and **E3**; $N = 91$). The experiments that asked participants to review their ballots saw significantly more time spent between ballot printing and submission (two-sample permutation test $p < 0.004$, 10k iterations), an average of 83 s (median 72 s) compared to 50 s without (median 33 s).

Notably, participants who were given a slate of candidates to vote for had much higher submission times (two-sample permutation test $p < 0.004$, 10k iterations). Noticers in the slate experiments took an average of 119 s (median 111 s) and non-noticers averaged 55 s (median 52 s). This might be partly attributed to voters having to select unfamiliar candidates and wanting to check their work.

[2]All *p*-values were computed with a Bonferroni correction at a family-wise error rate of 0.05.

3) Demographics: Comparisons of detection rates across demographic groups revealed that a strong indicator for verification performance was voting experience. Subjects who reported being registered to vote ($N = 220$) detected errors with their ballots 19% of the time, while their those who did not ($N = 21$) detected errors 4.8% of the time. Those who reported voting previously ($N = 216$) caught ballot discrepancies in 19% of cases, again performing better than those who reported not voting before ($N = 25$), who detected an error in 4.0% of cases. If someone reported voting in the 2018 midterm election ($N = 209$), they detected problems with their ballot 20% of the time, whereas if they did not ($N = 32$), they detected problems 3.1% of the time. This may indicate that familiarity with the midterm ballot we used caused participants to feel more invested in the accuracy of their votes; however, we did not establish this to statistical significance.

Other demographic factors, such as age, education, ethnicity, and gender, had no correlation with detecting manipulation.

4) Ballot position: Noticing was correlated with ballot position (Pearson's of -0.64), indicating that discrepancies in more prominent races are more likely to be noticed. (Race 0 was the first race on the ballot, so the number of noticers decreases as the race position increases, hence the negative correlation coefficient.) On our ballot, the first five races (Governor, Secretary of State, Attorney General, U.S. Senator, and Representative in Congress) were prominent partisan contests with a high likelihood of name recognition. In the experiments with no intervention (**E1–E3**), 37 participants had one of these races manipulated, and five reported the error on the exit survey, a rate of 14%. Additional experiments are necessary to establish the strength of this effect when combined with interventions.

5) Undervotes: A metric that may inform voters' ability and willingness to verify their ballot is how much care they take in filling out the ballot. There are two metrics we use to examine this: whether a participant voted in every contest on the ballot, and whether the participant voted in every available position on the ballot (e.g., in a vote-for-two contests, the participant selected two choices). Table II shows the rates of voting in every race and every position on the ballot, with **E8** and **E9** removed as they directed participants to vote in every position. Voters who noticed discrepancies voted in every race or every position at a higher rate than those who did not, but not significantly so (likely due to our small sample size). Since these undervotes are visible to malware running on a BMD, this correlation could be exploited by an attacker to focus cheating on voters who are less likely to carefully verify, provided future work more firmly establishes this link.

	Overall	Noticers	Non-noticers
Every race	64.3%	73.9%	63.0%
Every position	43.0%	47.8%	42.4%

TABLE II: *Participant Attentiveness.* Voters who noticed the discrepancy tended to vote in every race and ballot position more often than those who did not.

6) Partisanship: To assess the role partisanship plays in detection rates, we scored each ballot with a partisanship score, where a vote for a Democratic candidate was scored -1 and a vote for a Republican candidate was scored 1, and we take the absolute value of the sum. There were 11 opportunities to vote in a partisan way, so a participant who voted straight-party for either major party would achieve a score of 11. Excluding **E8** and **E9**, where voters were directed how to vote, the mean partisanship score for our participants was 8.3, and the median was 11. Although our BMD did not offer an automatic "straight-party" voting option, 105 participants achieved the maximum partisanship score.

Intuitively, a voter expecting every selected candidate to be from the same party might be more likely to notice a selection from a different party. Looking at only these straight-party voters, 15 out of 105 detected the errors. Of those, nine had a partisan race swapped to a different candidate of a different party, and six of those participants wrote in the survey that they had detected the change based on party. For example, one participant wrote, *"voted GOP for governor / lieutenant governor but Libertarian was actually selected on the paper ballot."*

This suggests that choosing a uniform set of candidates may help voters detect when something has gone wrong on their ballot, although more work is needed to establish that this is indeed the case, especially in more politically diverse populations. If this positive effect holds, it could be further promoted with ballot designs that prominently display the party, which could help voters see the information that is important to them while they review the ballot. On the other hand, BMD malware could be designed to counter this effect by focusing cheating on voters who do not cast a straight-party ballot.

7) Slate voting: 34 participants were assigned an intervention which asked them to vote for a preselected slate of candidates (with a partisanship score of 0). Of these, only 26 participants voted exactly as directed. Of the eight participants who did not, four voted a straight Democratic ticket (partisanship score of 11), one voted a heavily Democratic ticket (score of 9), two voted slightly Democratic tickets (scores of 3 and 5), and one voted a non-partisan ticket (score of 0), which only deviated from the slate in five positions. Of the eight participants who deviated from the slate, no participant deviated by fewer than five positions, indicating that either the deviation was deliberate or our instructions to vote the slate were unclear. Only one deviating participant managed to notice the discrepancy on their ballot, leaving participants who deviated from the slate a 13% notice rate compared to the 73% notice rate for those who did not deviate.

8) Network effects: One potential feature of a live polling place environment is a network effect: will a voter who is voting at the same time as a noticer be more likely to notice a problem on theirs? However, the number of people who notice in a given experiment is a confounding factor: voters are more likely to overlap with a noticer if there are more noticers. To interrogate this, we ran partial hypothesis tests for each intervention using Fisher's exact tests with permutations of overlapping with a noticer and noticing, and then combined using Fisher's combining function. We found that the effect of overlapping with a noticer did not significantly impact whether a participant noticed. This suggests that our interventions were more important than overlapping.

9) Signage: One feature that did not correlate with improved verification performance was the signage we tested (**E4**). Our observer noted that 11 of 30 participants in the signage experiment did not notice the sign at all. Only two participants in this experiment detected the modification of their ballot and reported it, and only one accurately noted the discrepancy in their survey, suggesting that passive signage alone may be insufficient to capture voters' attention and shape their subsequent behavior.

D. Participant Comments

Participants had two free-response sections in the exit survey. The first asked about anything "odd" they had noticed about the ballot. The second invited any additional comments. Of the 241 participants, 114 responded to at least one of these prompts. We note several features of their responses.

1) Discrepancy reports: In total, 44 participants (18%) noted in the free response section of the survey that they had identified some discrepancy on their paper ballot. Of these, 31 correctly identified the change, 12 gave no detail (e.g., *"At least one of my choices did not match who I picked"*), and one incorrectly identified the change (but did report that there was a mistake). We omitted this last participant from our "noticers" category where applicable.

Of the 44 participants who reported a change on their ballot in the survey, five added that they thought it could have resulted from a mistake they made. For example, one participant reported: *"I don't remember voting for the member of Congress and there was a vote. I very well may have but just don't remember."*

2) Attitudes about verification: Twelve participants mentioned either that they would only be comfortable voting on a paper ballot or that they were comforted by the fact that a paper trail was created. Only three of these 12 participants noticed that their ballot had been modified, despite the fact that they recognized that the paper ballot was an important tool for ensuring election integrity.

Several participants seemed to realize *after* casting their vote that the evaluation of their paper ballot was important; 13 participants mentioned in the survey that they did not review or that they should have reviewed the ballot, although we did not ask them about it. This concern may have been triggered by our survey question about what they had noticed about the paper ballot, but it also might be an indication that our interventions did cause voters to think about the risk—albeit too late.

The free responses also indicate that some participants assumed that the vote was completed and submitted on the BMD, rather than the paper ballot being the official record of their vote. One participant wrote, *"I was surprised to still have a paper ballot, after using the touch system. I was expecting the results to be registered electronically."* This assumption may discourage voters from verifying the selections on their

paper ballot. Similarly, another participant, prompted by script variant 3 ("Have you carefully reviewed each selection on your printed ballot?"), responded to a poll worker, *"I checked it on the screen, it better be right."*

Three participants expressed concern that they would not know what to do if they noticed a problem with their paper ballot during a real election. One person wrote, *"Having the printout be incorrect was confusing and it's not clear how that would be handled in an election environment."*

3) Feedback on the BMDs: We told participants that the experiment was a study about a new kind of voting system, and many left feedback about the interface and appearance of the machines. In Michigan, where we conducted the study, BMDs are available in every precinct, but voters must request to use them. The vast majority of voters use hand-marked paper ballots, so study participants were likely unfamiliar with BMD voting. In their comments, 21 participants expressed liking the system, while only three disliked it. Although merely anecdotal, this reflects previous findings that voters like touch-screen voting equipment [22].

V. SECURITY MODEL

We are primarily motivated by the threat of undetected changes to election outcomes due to BMD misprinting attacks. Prior work has shown that such attacks cannot be reliably ruled out by pre-election or parallel testing [51], and we seek to answer whether voter verification can be an effective defense.

If a voter reports that their printed ballot does not reflect their on-screen selections, what should election officials do? Unfortunately, there is not yet a practical way to prove that the BMD misbehaved during voting. From officials' perspective, it is also possible that the voter is mistaken, or even lying, and in a large voter population, there will always be some rate of spurious problem reports, even when BMDs are working correctly.

For these reasons, problem reports from voters can serve only as evidence that something *might* be wrong with the BMDs. If the evidence exceeds some threshold, officials could invoke contingency plans. For instance, they could remove BMDs from service to minimize further damage, perform forensic investigations in an attempt to uncover the cause, or even rerun the election if outcome-changing fraud cannot be ruled out.

Any of these responses would be costly (and none is foolproof), so the threshold for triggering them should not be too low. Moreover, attackers could exploit a low threshold by recruiting voters to fraudulently report problems, in order to disrupt or discredit the election. On the other hand, if the threshold is too high, outcome-changing fraud could be ignored.

To better understand how verification performance affects security in this setting, we construct a simple model. We assume, optimistically, that the attacker has no way to guess whether a particular voter is more likely than average to detect the alteration, and so chooses voters to attack at random. We further assume that whenever voters detect problems, they are able to remedy them and cast a correct vote by hand-marking a ballot. Except where noted, the model assumes that all voters cast their votes using BMDs.

Number of problem reports   Let d be the fraction of misprinted ballots that voters detect, report, and correct. Suppose a contest had n ballots cast, and the reported fractional margin of victory was m. To have changed the outcome, the attacker would have had to successfully modify at least $n\frac{m}{2}$ cast ballots. However, since some modifications would have been corrected, the attacker would have had to induce errors in a greater number of printouts: $n\frac{m}{2(1-d)}$. Under our optimistic assumptions, if the attack changed the outcome, we would expect the fraction of voters who reported problems, a, to exceed:

$$a > m\frac{d}{2(1-d)}.$$

The model shows that the security impact of verification is non-linear, because every voter who corrects an error *both* increases the evidence that there is a problem *and* forces the attacker to cheat more in order to overcome the margin of victory. Figure 4 illustrates this effect.

With the 6.6% error detection rate from our non-intervention experiments and a close election with a 0.5% margin (the margin that causes an automatic recount in many states) a successful attack would cause as few as 0.018% of voters—less than 1 in 5000—to report a problem. Small changes in verification performance around our base rate cause relatively little change in the amount of evidence. More than doubling the error detection rate to 14% (the rate we found for prominent races) only increases the fraction of voters who report a problem to 0.039%. However, larger improvements have an outsized effect: with the 86% error detection rate from our most successful experiment, at least 1.5% of voters (1 in 67) would report problems.

Required detection rate   Suppose election officials activate a countermeasure if the fraction of voters who report problems
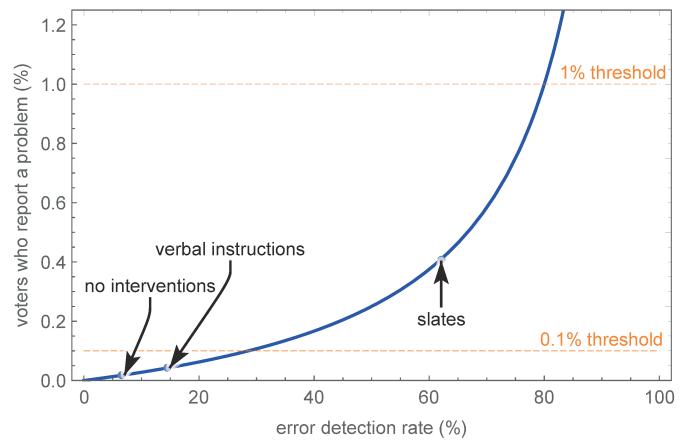

Fig. 4: *BMD security is highly sensitive to human performance.* Given a 0.5% margin of victory, we plot the percentage of voters who report a problem during the minimal outcome-changing attack as a function of the rate at which errors are detected and corrected. This model implies that using BMDs safely for all voters requires dramatically improved verification performance or very sensitive attack detection thresholds.

exceeds a threshold a^*. For a given margin, the countermeasure will be triggered by minimal outcome-changing fraud when:

$$d > \frac{2a^*}{m + 2a^*}.$$

An expensive countermeasure, like rerunning an election, will require a high trigger threshold—say, 1% of voters reporting a problem—to avoid false positives. With a 0.5% margin, reaching a 1% trigger threshold would require an error detection rate exceeding 80%. A less expensive countermeasure, such as an investigation, might be triggered by a lower threshold—say, 0.1%. Reaching this lower threshold in an election with a 0.5% margin would require an error detection rate greater than 29%. This suggests that using BMDs securely for all voters will require large improvements to verification performance or extremely low thresholds for triggering countermeasures.

Minimizing BMD voting helps dramatically   Securing against misprinting attacks is far easier if only a small fraction of voters use BMDs than if all in-person voters do. This is because an attacker would be forced to cheat on a much larger fraction of BMD ballots in order to achieve the same change to the election results. Moreover, if the population of BMD voters is smaller than half the margin of victory, it is impossible for a BMD misprinting attack to change the outcome.

Let b be the fraction of voters who use BMDs. We can replace m in the expression above with $\frac{m}{b}$ and let a^* be the fraction of *BMD voters* that must report a problem to trigger the countermeasure. In Maryland, which uses hand-marked paper ballots but makes BMDs available to voters who request them, 1.8% of voters use BMDs [34]. With a 0.5% margin, as in the previous example, Maryland would reach a complaint threshold of 1% of BMD voters with an error detection rate of only 6.7%. If 5% of voters use BMDs, the error detection rate would need to be 17%. Our results suggest that these more modest rates of verification likely are achievable, in contrast to the far greater accuracy required when all voters use BMDs.

This model overestimates security   An attacker might use any number of features (including several of the correlations we observed) to focus cheating on voters who are less likely to successfully catch errors. For instance, an attacker could preferentially modify ballots that have undervotes or a mix of selections from different parties. Attackers could also selectively target voters with visual impairments, such as those who use large text or an audio ballot. Other features, such as how long voters spend inspecting the candidate review screen, might also prove to be predictive of verification success. For these reasons, our simplified model is likely to overestimate the effectiveness of verification against sophisticated attackers.

We also note that some attackers may merely seek to cast doubt on election results by causing highly visible errors or failures—which are also possible with hand-marked paper ballots. However, in general, BMDs are vulnerable to all classes of computer-based attacks that affect hand-marked paper ballots and to others, such as the misprinting attack discussed here, to which hand-marked paper ballots are not susceptible.

VI. DISCUSSION

A. Limitations

It is challenging to capture real-world voter behavior in a mock election. However, our study followed established best practices [39], and we strived to create as realistic a polling environment as we could. It is impossible to know exactly how well we succeeded, but the effect seems to have been convincing: several people approached us to ask whether there was a real election taking place that they had not heard about. Our participants also seemed engaged in the study; many expressed strongly held political preferences in our survey (so much so that some refused to vote according to our slate), and a large majority reported voting in the 2018 midterm. On the other hand, the election used a ballot that was more than nine months old, which may have reduced participant motivation, and we had a few participants who reported that they did not vote in our state or were otherwise unfamiliar with our ballot. It is also possible that our results were skewed due to selection bias and observer effect.

Another limitation of our work is that we drew participants from a population that is locally but not nationally representative. Our participants tended to be younger, significantly better educated, more liberal, more likely to be female, and more likely to be Caucasian than the average voter in the United States [54]. Future work is needed to validate our study in more diverse and representative populations.

Although our results suggest that certain interventions can boost verification performance, the data is too sparse to provide a high-fidelity understanding of the magnitude of the improvements. In addition, due to time constraints, we were unable to test the interplay of all combinations of interventions, and some interventions appear to be sensitive to small changes (e.g., the difference in phrasing between script variants 2 and 3). Further study is needed to better characterize what makes interventions work and how they interact before we can confidently conclude that any particular set of procedures will be effective in practice.

B. Discussion of Findings

Our study provides the first concrete measurements of voter error detection performance using BMDs in a realistic voting environment. At a high level, we found that success rates without intervention are very low, around 6.6%. Some interventions that we tested did not significantly impact detection rates among participants, although others improved detection drastically and may serve as a roadmap for interventions to explore in further research. We discuss those interventions here.

1) Verbal instructions can improve verification: Notably, all interventions that involved poll workers verbally encouraging verification between the BMD and the scanner—those in **E6–E9**—resulted in higher ballot reviewing and error reporting rates. This, coupled with the fact that reviewing the printout was highly correlated with error detection across all of our results, suggests that interventions focused on causing the voter to review the ballot carefully may be helpful. On the

other hand, instructions at the beginning of the voting process (**E5**) and passive signage (**E4**) had no significant effect on error reporting. This pattern of effects is supported by findings from the usable security literature, which suggest that post-completion errors can be mitigated with timely interruptions that encourage individuals to take defensive steps [14].

It is worth noting that we also found that these interventions caused participants to take longer to submit their ballots, on average about twice as long. This could cause longer lines at polling places if these interventions are implemented without complementary procedural considerations, such as having adequate space for voters to stop and review their ballots.

2) Effectiveness of slates: Directing participants to vote for a provided slate of candidates, combined with verbally prompting them to review their printouts, resulted in strongly increased rate of error detection: 74% of participants who were given a slate and did not deviate from it noticed the errors. This finding may suggest that encouraging voters to write down their preferences in advance can boost verification.

However, the slates we used functioned quite differently from slates likely to be used in practice. The choices we provided were randomly generated and had no basis in the subject's preferences — in a real election, slates would reflect who the voter intended to vote for, most likely created by the voter or their political party [29]. It is possible that the success rate we observed was primarily due to participants carefully attempting to follow our instructions and vote for unfamiliar candidates. Further study is needed with more realistic slate conditions (i.e., asking subjects to write down their preferences) in order to assess whether slates really do help voters catch errors.

C. Recommendations

Since BMDs are widely used today, we recommend several strategies for improving voter verification performance. While we are unable to conclude that these strategies will enhance error detection to the point that BMDs can be used safely in close or small elections, our findings indicate that they can help.

1) Design polling places for verification: Polling place layout and procedures should be designed with verification in mind. As we have discussed, voters need time and space to verify their ballots. If tables or areas to stand out of the way are provided, voters will be able to carefully verify without causing lines to form or slowing polling place throughput. The presence of such a "verification station" might also encourage verification.

Another practical concern is privacy. Several of our participants expressed discomfort with the fact that we did not provide a privacy sleeve for their ballots (a requirement in Michigan), and that the scanner accepted the ballots face-up only, with one participant stating, *"I feel like inserting the ballot face up in the scanning machine will make people uncomfortable."* Voters may not feel comfortable reviewing their ballots in front of poll workers but may be unsure where to go to review them privately.

2) Incorporate post-voting verbal instructions: As all of our script-based interventions that took place after the ballot was printed (**E6–E9**) showed an increase in verification performance, we recommend that poll workers interrupt voters after their ballot has printed but before it is scanned and ask them to review it. Signage with a similar message to our scripts placed at the optical scanner (**E4**) or instructions before the participants voted (**E5**) did not result in significant differences in error detection; nevertheless, further study with additional variations is prudent before ruling out such strategies.

3) Encourage personalized slate voting: Although our study tested randomized slates, rather than personalized slates, the effect size was so large that we tentatively recommend encouraging the use of personalized slates by voters. In our experiments (**E8** and **E9**), participants who were directed to vote using a randomized slate (and did not deviate) reported errors at a rate of 73%. If voters prepare their own slates at home (or use a printed slate prepared, for instance, by a political party or other organization), they can use them to check each selection on the BMD printout. We note that, since we did not directly test the use of personalized slates, further research is necessary to ascertain whether large performance gains are actually achieved. Furthermore, even if personalized slates are effective, the gain will be limited to the fraction of voters who can be induced to use them.

Slates have potential downsides and should be used with care. They have the potential to compromise ballot secrecy, so we recommend providing a closed trash can, paper shredder, or other means for voters to privately dispose of them before leaving the precinct. Coercion is also a threat, but voters could be advised to prepare multiple different slates as a defense.

4) Help voters correct errors, and carefully track problems: Verification-promoting interventions will be of little use if action cannot be taken to remedy misbehaving BMDs — something that even our participants expressed concern about.

First, it is crucial that polling places have a procedure for voters who want to correct their printed ballots. Several subjects commented that they would not know what to do if something was wrong with their ballot in a real election, indicating that this problem is present in current election procedures.

Second, detailed records should be kept about which BMD the voter used and what the specific issue was, including the contest and candidates involved (to the extent that the voter is willing to waive ballot secrecy). Problems should be treated as potentially serious even when the voter believes they are at fault — we note that several participants in our study believed they had made a mistake even though the BMD actually was programmed to be malicious. Problem reports should be centrally reported and tracked during the election, so that issues affecting multiple precincts can be identified as rapidly as possible.

5) Prepare contingency plans: What to do in the event that BMDs are known or suspected to be misbehaving is a more difficult question. If an elevated number of voters have a problem with a single machine, it should be taken out of service, provided there are other BMDs available for use (especially for voters with disabilities, who may have no alternative).

If widespread problem reports occur — particularly problems focused on a tightly contested race or significantly exceeding the rate reported in past elections — officials could consider

taking most BMDs out of service and encouraging all remaining voters who can to use hand-marked ballots. This raises logistical challenges: polling place would need to have enough ballots available for hand-marking, or the ability to print ballots on demand, and votes already cast on the BMDs would be suspect.

After the election, forensic analysis of the BMDs could be performed to attempt to determine the cause of reported errors. Unfortunately, such analysis cannot in general rule out that a sophisticated attack occurred and left no digital traces. Even if programming errors or attacks are uncovered, they may be impossible to correct if officials are unable to determine whether the effects were large enough to change the election outcome. The only recourse might be to re-run the election.

Our findings show that, in the event of an actual error or attack, the rate of reported problems is likely to be only the tip of the iceberg. In our non-intervention experiments, undetected errors outnumbered reported problems by almost twenty to one. Our results further suggest that an attacker who cleverly focused cheating on voters who were less likely to verify could achieve an even higher ratio of undetected errors. An effective response requires either being very sensitive to reported problems—which increases the chances that an attacker could trigger false alarms—or achieving very high error correction rates.

6) Educate voters about BMD operations and risks: Like in other human-in-the-loop security contexts, greater education could boost voters' awareness of the importance of careful verification and boost error detection and reporting rates.

To this end, we recommend educating voters that the paper, rather than what the BMD screen shows, is the official record of their votes. Several of our participants said they realized after scanning that they should have, but did not, review their printouts. Others stated that they had checked the review screen on the machine and that they trusted the paper to be correct. It is likely that many participants incorrectly assumed that the BMDs, rather than the paper and scanner, tabulated their votes.

We also recommend educating voters about the possibility of BMD malfunction. Many of our participants seem not to have even considered that the machine might have changed their votes, as indicated by the voters who blamed themselves for the misprinted ballots. Raising threat awareness could help motivate voters to carefully inspect the paper, as well as give them greater confidence to report any discrepancies they detect.

7) Consider the needs of voters with disabilities: Further research is needed to specifically examine verification performance among voters with disabilities, but we offer some initial recommendations here. Detecting errors in printed ballots may be especially challenging for voters with impaired vision. Designing BMD ballots for maximum legibility might help, and so might encouraging voters who use text-to-speech devices to bring them to the polls for use during verification. Jurisdictions could also provide air-gapped accessible devices to read the ballot back to voters, in case voters do not have their own text-to-speech devices. These steps would have the added benefit of reinforcing the message that the content of the paper ballots is what gets counted. If BMDs are to live up to the promise of

better and more accessible voting, enabling all voters to verify their printed ballots is a must.

8) Require risk-limiting audits: Even perfectly verified paper ballots are of little use for security if they are not rigorously audited to confirm the results of computer-based tabulation. Fortunately, risk-limiting audits [32] (RLAs) are gaining momentum in the United States. Colorado, Nevada, and Rhode Island mandate statewide RLAs, and states including Michigan, Virginia, Georgia, and Pennsylvania are considering implementing them soon [17]. RLAs and effective verification are both necessary in order for paper to provide a strong defense against vote-stealing attacks, and we recommend that efforts to achieve both be pursued vigorously.

VII. CONCLUSION

We conducted the first empirical study of how well voters using BMDs detect errors on their printed ballots, which is a limiting factor to the level of security that a BMD-based paper trail can provide. Based on the performance of 241 human subjects in a realistic polling place environment, we find that, absent specific interventions, error detection and reporting rates are dangerously low. Unless verification performance can be improved dramatically, BMD paper trails, particularly when used by all in-person voters, cannot be relied on to reflect voter intent if the machines are controlled by an attacker.

Nevertheless, we also find that procedural interventions can improve rates of error detection and reporting, potentially increasing the security offered by BMDs. The interventions we tested should serve as examples of what is and is not likely to be effective, and we hope they will point the way for further research and experimentation. These findings add to the broad literature of human-in-the-loop security results and recommendations, and they provide additional examples of what does and does not work in human-centric security.

Our results should not be read as demonstrating that BMDs can be used securely. Further work is needed to explore the potential for attackers to predict which voters will verify, and additional human-subjects testing is necessary to confirm whether sufficient rates of verification success can be achieved in practice. The cost of implementing interventions and contingency plans may also be prohibitive. Nevertheless, BMDs do offer advantages, including uniform accessibility and ease of administration. We hope our work will help election officials make better informed choices as they weigh these benefits against the security risks of using BMDs for all voters.

REFERENCES

[1] C. Z. Acemyan, P. Kortum, and D. Payne. Do voters really fail to detect changes to their ballots? An investigation of ballot type on voter error detection. *Proceedings of the Human Factors and Ergonomics Society*, 57:1405–1409, 2013.

[2] D. Akhawe and A. P. Felt. Alice in Warningland: A large-scale field study of browser security warning effectiveness. In *22nd USENIX Security Symposium*, pages 257–272, 2013.

[3] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, and Y. Agarwal. Your location has been shared 5,398 times! A field study on mobile app privacy nudging. In *33rd ACM Conference on Human Factors in Computing Systems*, CHI, pages 787–796, 2015.

[4] Ann Arbor District Library. Classic shop drop! Plus, fish election results!, Aug. 2019. https://aadl.org/node/396262.

[5] Ann Arbor District Library. Mock the vote, July 2019. https://aadl.org/node/395686.

[6] Ann Arbor District Library. Mock voting @ AADL, Sept. 2019. https://aadl.org/node/397364.

[7] A. Appel, R. DeMillo, and P. Stark. Ballot-marking devices (BMDs) cannot assure the will of the voters, 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3375755.

[8] S. Bell, J. Benaloh, M. D. Byrne, D. DeBeauvoir, B. Eakin, G. Fisher, P. Kortum, N. McBurnett, J. Montoya, M. Parker, O. Pereira, P. B. Stark, D. S. Wallach, and M. Winn. STAR-Vote: A secure, transparent, auditable, and reliable voting system. *USENIX Journal of Election Technology and Systems*, 1(1), 2013.

[9] M. Bernhard, J. Benaloh, J. A. Halderman, R. L. Rivest, P. Y. Ryan, P. B. Stark, V. Teague, P. L. Vora, and D. S. Wallach. Public evidence from secret ballots. In *2nd International Joint Conference on Electronic Voting*, E-Vote-ID, pages 84–109, 2017.

[10] R. Bernstein, A. Chadha, and R. Montjoy. Overreporting voting: Why it happens and why it matters. *Public Opinion Quarterly*, 65(1):22–44, 2001.

[11] D. Bowen et al. Top-to-bottom review of voting machines certified for use in California. Technical report, California Secretary of State, 2007. https://www.sos.ca.gov/elections/ovsta/frequently-requested-information/top-bottom-review/.

[12] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri. Bridging the gap in computer security warnings: A mental model approach. *IEEE Security & Privacy*, 9(2):18–26, Mar. 2011.

[13] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter. Your attention please: Designing security-decision UIs to make genuine risks harder to ignore. In *9th Symposium on Usable Privacy and Security*, SOUPS, 2013.

[14] M. D. Byrne and S. Bovair. A working memory model of a common procedural error. *Cognitive Science*, 21(1):31–61, 1997.

[15] B. A. Campbell and M. D. Byrne. Now do voters notice review screen anomalies? A look at voting system usability. In *USENIX Electronic Voting Technology Workshop/Workshop on Trustworthy Elections*, EVT/WOTE, 2009.

[16] Clear Ballot Group. ClearAccess administrators guide, 2015. https://www.sos.state.co.us/pubs/elections/VotingSystems/systemsDocumentation/ClearBallot/ClearAccess/ClearAccessAdministratorsGuideRev4-0-r0.pdf.

[17] A. Cordova, L. Howard, and L. Norden. Voting machine security: Where we stand a few months before the New Hampshire primary. Brennan Center, 2019. https://www.brennancenter.org/analysis/voting-machine-security-where-we-stand-six-months-new-hampshire-primary.

[18] L. F. Cranor. A framework for reasoning about the human in the loop. In *1st Conference on Usability, Psychology, and Security*, UPSEC. USENIX, 2008.

[19] M. De Jong, J. Van Hoof, and J. Gosselt. Voters' perceptions of voting technology: Paper ballots versus voting machine with and without paper audit trail. *Social Science Computer Review*, 26(4):399–410, 2008.

[20] R. DeMillo, R. Kadel, and M. Marks. What voters are asked to verify affects ballot verification: A quantitative analysis of voters' memories of their ballots, 2018. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3292208.

[21] S. Egelman, L. F. Cranor, and J. Hong. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *26th ACM Conference on Human Factors in Computing Systems*, CHI, pages 1065–1074, 2008.

[22] S. P. Everett, K. K. Greene, M. D. Byrne, D. S. Wallach, K. Derr, D. Sandler, and T. Torous. Electronic voting machines versus traditional methods: improved preference, similar performance. In *26th ACM Conference on Human Factors in Computing Systems*, CHI, pages 883–892, 2008.

[23] A. J. Feldman, J. A. Halderman, and E. W. Felten. Security analysis of the Diebold AccuVote-TS voting machine. In *USENIX Electronic Voting Technology Workshop*, EVT, 2007.

[24] M. Fitts. SC chooses new voting machines that will print paper ballots but some fear it's not safe. The Post and Courier, June 10, 2019. https://www.postandcourier.com/article_f86632ce-8b83-11e9-8dab-5fb7858906cc.html.

[25] S. Fowler. Georgia awards new voting machine contract to Dominion Voting Systems. Georgia Public Broadcasting, July 29, 2019. https://www.gpbnews.org/post/georgia-awards-new-voting-machine-contract-dominion-voting-systems.

[26] S. N. Goggin and M. D. Byrne. An examination of the auditability of voter verified paper audit trail (VVPAT) ballots. In *USENIX Electronic Voting Technology Workshop*, EVT, 2007.

[27] K. K. Greene, M. D. Byrne, and S. P. Everett. A comparison of usability between voting methods. In *USENIX Electronic Voting Technology Workshop*, EVT, 2006.

[28] Indiana Fiscal Policy Institute. Vote centers and election costs: A study of the fiscal impact of vote centers in Indiana, 2010. https://www.in.gov/sos/elections/files/IFPI_Vote_Centers_and_Election_Costs_Report.pdf.

[29] D. Jones and B. Simons. *Broken Ballots: Will Your Vote Count?* CSLI Publications, 2012.

[30] D. Kasdan. Early voting: What works. https://www.brennancenter.org/sites/default/files/publications/VotingReport_Web.pdf.

[31] T. Kohno, A. Stubblefield, A. D. Rubin, and D. S. Wallach. Analysis of an electronic voting system. In *25th IEEE Symposium on Security and Privacy*, 2004.

[32] M. Lindeman and P. Stark. A gentle introduction to risk-limiting audits. *IEEE Security & Privacy*, 10:42–49, 2012.

[33] J. Mack. Who votes in Michigan? A demographic breakdown. MLive, 2018. https://www.mlive.com/news/erry-2018/11/340b0f9c406363/who-votes-in-michigan-a-demogr.html.

[34] S. Maneki and B. Jackson. Re: Comments on Ballot Marking Devices usage for the 2018 elections, 2017. Letter to Maryland State Board of Elections, citing SBE data.

[35] P. McDaniel, M. Blaze, and G. Vigna. EVEREST: Evaluation and validation of election-related equipment, standards and testing. Technical report, Ohio Secretary of State, 2007. https://www.eac.gov/assets/1/28/EVEREST.pdf.

[36] National Academies of Sciences, Engineering, and Medicine. *Securing the Vote: Protecting American Democracy*. The National Academies Press, Washington, DC, 2018.

[37] National Conference of State Legislatures. Funding elections technology, 2019. https://www.ncsl.org/research/elections-and-campaigns/funding-election-technology.aspx.

[38] R. G. Niemi and P. S. Herrnson. Beyond the butterfly: The complexity of U.S. ballots. *Perspectives on Politics*, 1(2):317–326, 2003.

[39] M. M. Olembo and M. Volkamer. E-voting system usability: Lessons for interface design, user studies, and usability criteria. In *Human-Centered System Design for Electronic Governance*, pages 172–201. IGI Global, 2013.

[40] S. Patil, R. Hoyle, R. Schlegel, A. Kapadia, and A. J. Lee. Interrupt now or inform later? Comparing immediate and delayed privacy feedback. In *33rd ACM Conference on Human Factors in Computing Systems*, CHI, pages 1415–1418, 2015.

[41] J. Petelka, Y. Zou, and F. Schaub. Put your warning where your link is: Improving and evaluating email phishing warnings. In *37th ACM Conference on Human Factors in Computing Systems*, CHI, 2019.

[42] Pew Charitable Trusts. Colorado voting reforms: Early results. https://www.pewtrusts.org/ /media/assets/2016/03/coloradovotingreformsearlyresults.pdf, 2016.

[43] Pro V&V. Test report for EAC 2005 VVSG certification testing: ClearBallot Group ClearVote 1.4 voting system, 2017. https://www.eac.gov/file.aspx?A=kOBM5qPeI8KZlJyADXYTieiXLwsxw4gYKIVroEkEBMo%3D.

[44] W. Quesenbery. Ballot marking devices make voting universal. Center for Civic Design, 2019. https://civicdesign.org/ballot-marking-devices-make-voting-universal/.

[45] W. Quesenbery, J. Cugini, D. Chisnell, B. Killam, and G. Reddish. Letter to the editor: Comments on "A methodology for testing voting systems". *Journal of Usability Studies*, 2(2):96–98, 2007.

[46] R. W. Reeder, A. P. Felt, S. Consolvo, N. Malkin, C. Thompson, and S. Egelman. An experience sampling study of user reactions to browser warnings in the field. In *36th ACM Conference on Human Factors in Computing Systems*, CHI, 2018.

[47] R. Rivest. On the notion of 'software independence' in voting systems. *Philos. Trans. Royal Soc. A*, 366(1881):3759–3767, October 2008.

[48] T. Selker, E. Rosenzweig, and A. Pandolfo. A methodology for testing voting systems. *Journal of Usability Studies*, 2(1):7–21, 2006.

[49] T. Selker, E. Rosenzweig, and A. Pandolfo. Reply to comment on: The Methodology for Testing Voting Systems by Whitney Quesenbery, John Cugini, Dana Chisnell, Bill Killam, and Ginny Redish. *Journal of Usability Studies*, 2(2):99–101, 2007.

[50] P. Stark. Conservative statistical post-election audits. *Annals of Applied Statistics*, 2(2):550–581, 2008.

[51] P. B. Stark. There is no reliable way to detect hacked ballot-marking devices, 2019. https://arxiv.org/abs/1908.08144.

[52] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying wolf: An empirical study of SSL warning effectiveness. In *18th USENIX Security Symposium*, pages 399–416, 2009.

[53] United States Senate Select Committee on Intelligence. Report on Russian active measures campaigns and interference in the 2016 U.S. election, 2019. https://intelligence.senate.gov/sites/default/files/documents/Report_Volume1.pdf.

[54] U.S. Census Bureau. QuickFacts: Ann Arbor, 2019. https://www.census.gov/quickfacts/annarborcitymichigan.

[55] U.S. Election Assistance Commission. Designing polling place materials. https://www.eac.gov/election-officials/designing-polling-place-materials/.

[56] Verified Voting. Ballot marking devices. https://www.verifiedvoting.org/ballot-marking-devices/.

[57] Verified Voting. The Verifier: Polling place equipment. https://www.verifiedvoting.org/verifier/.

[58] VSAP. Voting system for all people. https://vsap.lavote.net/.

[59] Wall Street Journal. Election 2018: How we voted in the 2018 midterms, November 6, 2018. https://www.wsj.com/graphics/election-2018-votecast-poll/.

[60] D. S. Wallach. On the security of ballot marking devices, 2019. https://arxiv.org/abs/1908.01897.

[61] M. S. Wogalter. Communication-human information processing (C-HIP) model. In M. S. Wogalter, editor, *Handbook of Warnings*, chapter 5, pages 51–61. Lawrence Erlbaum Associates, Mahwah, NJ, 2006.

[62] M. S. Wogalter and K. R. Laughery. Warning! sign and label effectiveness. *Current Directions in Psychological Science*, 5(2):33–37, 1996.

Our poll workers followed four versions of the script below: a baseline version, and three variants that each add one line.

VARIANT 1: Before the voter begins using the BMD, a poll worker asks them to check their ballot before it is scanned.

VARIANT 2: Before the voter deposits the ballot, a poll worker informs them that it is the official record of the vote.

VARIANT 3: Before the voter deposits the ballot, a poll worker asks whether they have carefully reviewed each selection.

When Subject Arrives (POLL WORKER A)

Hello! Before you begin, please fill out this Institutional Review Board consent form. [Point to form and pen.] If you have any questions, feel free to ask.

You are about to participate in a study about the usability of a new type of voting machine. You will be using one of these voting machines to make selections on your ballot, which will be a truncated version of the Ann Arbor 2018 midterm ballot. Once you are finished, your ballot will be printed from the printer beneath the machine, and you can review your ballot and deposit it in the ballot box over there. [Point out ballot box.] Feel free to vote your political preference or not; no identifying information will be collected that could match you with your votes. If you would like to quit at any time during the study, just say so.

VARIANT 1*: Please remember to check your ballot carefully before depositing it into the scanner.*

You may begin at any time.

Before Subject Deposits Ballot (POLL WORKER B)

VARIANT 2*: Please keep in mind that the paper ballot is the official record of your vote.*

VARIANT 3*: Have you carefully reviewed each selection on your printed ballot?*

After Subject Deposits Ballot (POLL WORKER B)

Thank you for participating! You are now finished with the study, and should fill out the exit survey. [Point to debrief survey computers.]

After Subject Completes Exit Survey (POLL WORKER B)

Thank you for your participation! You are now finished. If you have any questions about this study, you may ask them now, although I am unable to answer some questions due to the nature of the research. Here is a debrief form. [Hand subject a debrief form.] If you think of anything after you leave, you can reach [me/the principle investigators] through the information on the debrief form.

If you know anyone who might like to participate, please refer them here; we will be here [remaining time].

Thank you again for participating!

An investigator used the following script to recruit library patrons to participate in the study:

Hello, do you have 10 minutes to participate in a study about a new kind of voting machine that is used in elections across the United States? This study will consist of voting using our voting machine and depositing a printed paper ballot into a ballot box, and then filling out a survey about the experience. If you would like to participate, we will need you to first sign a consent form. We will provide a flyer at the end of your participation with information about the study. We cannot make all details available at this time, but full details and research results will be made available within six months of the conclusion of this study. We thank you for your consideration and hope you choose to participate!

We randomly generated a slate of candidates and provided a printed copy to voters in certain experiments. The handout voters received is reproduced below:

Race	Candidate(s)
Governor and Lieutenant Governor	Bill Gelineau and Angelique Chaiser Thomas
Secretary of State	Mary Treder Lang
Attorney General	Lisa Lane Gioia
United States Senator	Debbie Stabenow
Representative in Congress 12th District	Jeff Jones
Member of State Board of Education (Vote for 2)	Tiffany Tilley, Mary Anne Hering
Regent of the University of MIchigan (Vote for 2)	Jordan Acker, Joe Sanger
Trustee of Michigan State University (Vote for 2)	Mike Miller, Bruce Campbell
Justice of the Supreme Court (Vote for 2)	Megan Kathleen Cavanagh, Kerry Lee Morgan
Judge of Court of Appeals 3rd District Incumbent Position (Vote for 2)	Jane Marie Beckering, Douglas B. Shapiro
Judge of Circuit Court 22nd Circuit Incumbent Position (Vote for 2)	Timothy Patrick Connors, Carol Kuhnke
Judge of Probate Court Incumbent Position	Darlene A. O'Brien
Judge of District Court 14A District Incumbent Position	Thomas B. Bourque