Testing Cannot Tell Whether Ballot-Marking Devices Alter Election Outcomes

Def Con Voting Village

Philip B. Stark and Ran Xie

7 August 2020

University of California, Berkeley

Why test BMDs?

- BMDs can print votes that differ from those confirmed onscreen or through audio interface.
- "voter-verifiability" & ability to spoil ballot don't solve the problem; c.f. Bernhard et al. 2020.
- In effect, BMDs make the paper trail hackable.

The BMD security model is broken

- BMDs make voters responsible for BMD security
- but BMDs don't give voters the tools they need to do that job
- no way for voter to prove BMD misbehaved
- LEO can't tell whether voter's complaint is BMD malfunction, voter error, or "wolf"
- error or malfeasance could change a large percentage of votes without raising an alarm

Claimed benefits of BMDs

- prevent overvotes
- warn of undervotes
- eliminate ambiguous marks

But . . .

- Assume BMDs function correctly!
  - Many recent examples of failures, including Georgia, Northampton PA, Los Angeles CA
- PCOS can also protect against undervotes and overvotes—required by VVSG 1.0

Can we establish that BMDs worked in a given election?

- need to know errors didn't change any outcomes
- 3 approaches proposed:
  - pre-election logic and accuracy (L&A) testing
  - "passive" testing
  - "live" or "parallel" testing
- Will show none of these can work in practice

How much testing is enough?

- depends on the size of the problem deemed "material."
- sensible threshold: "enough to change the reported winner of one or more contests"
- many contests are decided by less than 1%
- margin in the 2016 U.S. presidential election was 0.22% in MI, 0.37% in RI, 0.72% in PA, and 0.76% in WI

Auditing as an adversarial game

- Mallory seeks to alter the outcome of one or more contests in an election.
  - M does not want to be detected.
  - M knows the testing strategy
  - M knows the state history of each machine
  - M has a good model of voter behavior
- Pat seeks to ensure that any BMD problem that alters one or more outcomes will be detected.
  - P must obey the law and protect voter privacy.
  - P does not know which contest(s) M will attack nor M's strategy.

Jurisdiction sizes, contest sizes, margins

Important contests have sizes ranging from dozens of eligible voters to millions of eligible voters.

- median turnout in the 3017 U.S. counties in 2018 was 2,980 voters,
- less than 43,000 voters for more than 2/3 of jurisdictions
- In 73% of states, more than 50% of counties have fewer than 30,000 active voters.
- In 92% of states, $>$50% of counties have fewer than 100,000 active voters.
- in 2019, 317 U.S. cities had populations of 100,000 or more, out of over 19,500 incorporated places
  - if 80% of the population is of voting age & turnout is 55%, contests for elected officials in 98% of incorporated places involve fewer than $44,000$ voters.
- 2019 median population of U.S. incorporated areas is 725: ~50% of the 19,500 incorporated places have turnout $<=$320 voters.

2018 turnout by county

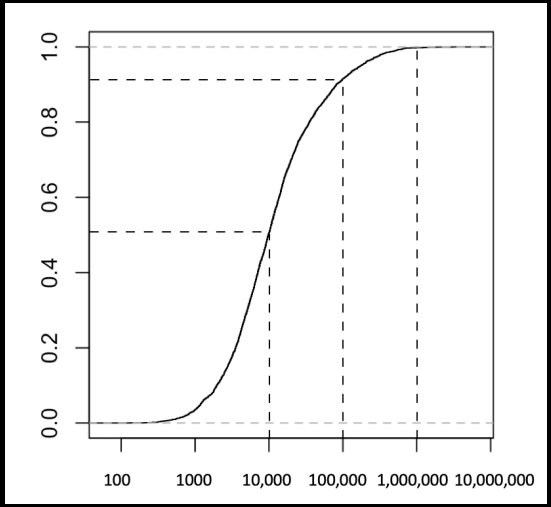

Figure 1: Total participation on election day per jurisdiction in 3073 counties in 2018 [@EAVS2018]. Counties ordered from small to large, plotted against total voter turnout.

2018 median turnout by jurisdiction

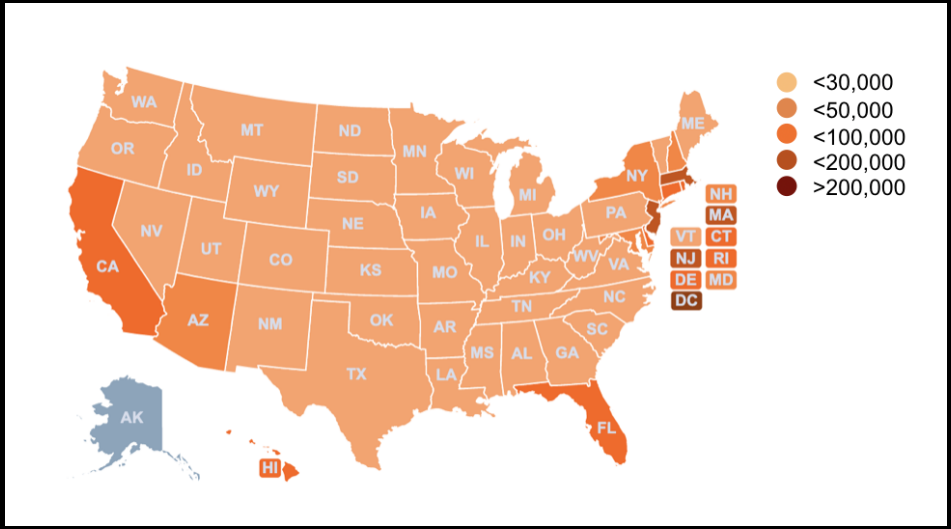

Figure 2: Heat map of median 2018 turnout by jurisdiction in the 50 U.S. states and

Mallory's strategy space: alter any collection of transactions

- time of day the transaction starts
- the time since the previous voter finished using the BMD (a measure of how busy the polling place is)
- the number of voting transactions before the current transaction
- the voter's sequence of selections in each contest, including undervotes, before going to the next selection
- the number of times the voter changes selections in each contest in the first pass through the ballot, and what the voter changed the selection from and to, etc.
- the amount of time the voter takes to make each selection before taking another action (e.g., going to the next contest)
- whether the voter looks every page of candidates in a contest

- how much time (if any) the voter takes to review selections, which selections the voter changes, etc.
- whether the voter receives an inactivity warning during voting
- what part of each onscreen voting target the voter touches
- BMD settings, including font size, language, whether the audio interface is used, volume setting, tempo setting, whether voter pauses the audio, whether voter "rewinds," and whether the voter uses audio only or synchronized audio/video
- whether voter uses sip-and-puff interface

Possible voting transactions

Parameter	optimistic	more realistic
Contests	3	20
Candidates per Contest	2	4
Languages	2	13
Time of day	10	20
Number of previous voters	5	10
Undervotes	2^3	2^{20}
Changed selections	2^3	2^{20}
Review	2	2
Time per selection	2	5^{20}

Parameter	optimistic	more realistic
Contrast/saturation	-	4
Font Size	2	4
Audio Use	2	2
Audio tempo	-	4
Volume	5	10
Audio pause	-	2^{20}
Audio + video	-	2
Inactivity warning	2	2^{20}
Total combinations	6.14×10^6	1.2×10^{47}

Pat's strategy space

- Monitor voter behavior, e.g., spoiled ballot rates

Pat's strategy space

- Monitor voter behavior, e.g., spoiled ballot rates
- Try to catch a malfunction by using the BMD before, during, or after an election

Randomness is key

- if Pat's tests are predictable, Mallory can just change other transactions (passive testing doesn't solve this)

Randomness is key

- if Pat's tests are predictable, Mallory can just change other transactions (passive testing doesn't solve this)
- can't just set aside machines: Dieselgate

Randomness is key

- if Pat's tests are predictable, Mallory can just change other transactions (passive testing doesn't solve this)
- can't just set aside machines: Dieselgate
- uniform random sampling is doomed

Randomness is key

- if Pat's tests are predictable, Mallory can just change other transactions (passive testing doesn't solve this)
- can't just set aside machines: Dieselgate
- uniform random sampling is doomed
- "ideal" sampling would mimic voter behavior

Randomness is key

- if Pat's tests are predictable, Mallory can just change other transactions (passive testing doesn't solve this)
- can't just set aside machines: Dieselgate
- uniform random sampling is doomed
- "ideal" sampling would mimic voter behavior
- examine "oracle bounds" and "learning" distribution of transactions

How many votes must be altered to alter the outcome?

- Altering votes on 1% of transactions in a jurisdiction can change the margin of contests that are not jurisdiction-wide by far more than 2%.

How many votes must be altered to alter the outcome?

- Altering votes on 1% of transactions in a jurisdiction can change the margin of contests that are not jurisdiction-wide by far more than 2%.
- If a contest is on 10% of ballots & undervote rate in the contest is 30%, altering votes on 1% of transactions can change margin in that particular contest by 29%.

Passive testing

- rely on voters to test
- use spoiled ballot rate to signal a possible problem
- need to set alarm threshold to balance false alarms and missed problems
- may depend on things that vary from election to election:
  - number of contests on the ballot
  - whether the contests have complex voting rules
  - ballot layout
  - voter demographics
  - turnout
  - familiarity w voting technology
  - . . .

Setting the threshold

- need to know something about the distribution of spoiled ballots when BMDs malfunction to control the false negative rate
- depends on the number of transactions Mallory alters, which voters are affected, which contests are affected, etc.
- Pat won't know any of those things

Hypothetical example

- spoiled ballots follow Poisson distribution with known rate, absent hacking, and different known rate, given hacking. (Optimistic!)
- 7% or 25% of voters will notice errors and spoil their ballots
- contest margins of 1%–5% and false positive and false negative rates of 5% and 1%.

5% false negative & false positive rate

margin	detection rate	0.5% base rate	1% base rate	1.5% base rate
1%	7%	451,411	893,176	1,334,897
	25%	37,334	71,911	106,627
2%	7%	115,150	225,706	336,160
	25%	9,919	18,667	27,325
3%	7%	52,310	101,382	150,471
	25%	4,651	8,588	12,445
4%	7%	30,000	57,575	85,227
	25%	2,788	4,960	7,144
5%	7%	19,573	37,245	54,932
	25%	1,838	3,274	4,689

1% false negative & false positive rate

margin	detection rate	0.5% base rate	1% base rate	1.5% base rate
1%	7%	908,590	1,792,330	2,675,912
	25%	76,077	145,501	214,845
2%	7%	233,261	454,295	675,242
	25%	20,624	38,039	55,442
3%	7%	106,411	204,651	302,864
	25%	9,870	17,674	25,359
4%	7%	61,385	116,631	171,908
	25%	5,971	10,312	14,681
5%	7%	40,156	75,671	110,989
	25%	4,036	6,849	9,650

Sanity check

- 41 of California's 58 counties had fewer than 100,000 voters in the 2018 midterm election

Sanity check

- 41 of California's 58 counties had fewer than 100,000 voters in the 2018 midterm election
- 33 had fewer than 100,000 voters in the 2016 Presidential primary election

Sanity check

- 41 of California's 58 counties had fewer than 100,000 voters in the 2018 midterm election
- 33 had fewer than 100,000 voters in the 2016 Presidential primary election
- passive testing could not protect contests with margins of 3% or smaller.

Sanity check

- 41 of California's 58 counties had fewer than 100,000 voters in the 2018 midterm election
- 33 had fewer than 100,000 voters in the 2016 Presidential primary election
- passive testing could not protect contests with margins of 3% or smaller.
- In many California counties, turnout is so small even in statewide contests that there would be no way to detect problems through spoilage rates reliably without high rate of false alarms.

Sanity check

- 41 of California's 58 counties had fewer than 100,000 voters in the 2018 midterm election
- 33 had fewer than 100,000 voters in the 2016 Presidential primary election
- passive testing could not protect contests with margins of 3% or smaller.
- In many California counties, turnout is so small even in statewide contests that there would be no way to detect problems through spoilage rates reliably without high rate of false alarms.
- If turnout is roughly 50%, contests in jurisdictions with fewer than 60,000 voters (which includes 23 of California's 58 counties could not in principle limit chance of false positives & of false negatives to 5% for margins below 4%—even under these optimistic assumptions and simplifications.

Targeting the attack

- assumed all voters are equally likely to detect discrepancies
- Mallory has access to each BMD's settings, state history, etc.
- can select whose votes to alter, inferring voter characteristics from BMD settings and the voters' interaction with the BMD.
- can target voters less likely to notice problems (&perhaps less likely to be believed if they report malfunctions)

Voters with visual impairments

- ~0.8% of the U.S. population is legally blind; approximately 2% of Americans age 16 to 64 have a visual impairment.
- Current BMDs do not provide voters with visual impairments a way to check whether the printout matches their selections
- If 2% of voters have a visual impairment that prevents them from checking the printout and Mallory only alters votes when the voter uses the audio interface or large fonts, Mallory might be able to change the outcomes of contests with jurisdiction-wide margins of 4% or more wo increasing the spoiled ballot rate.

Voters with motor impairments

- Some BMDs let voters print & cast a ballot without looking at it, e.g. ES&S ExpressVote® with "Autocast,"
- Voters who use this feature have no opportunity to check whether the printout matches their selections nor to spoil the ballot if there is a discrepancy.
- Mallory can change every vote cast using this feature without increasing the spoiled ballot rate.

Voters who use languages other than English

- Federal law requires some jurisdictions to provide ballots in languages other than English.
- In 2013, ~26% of voters in Los Angeles County spoke a language other than English at home
- If a substantial percentage of voters use foreign-language ballots and are unlikely to check the English-language printout, Mallory could change the outcome of contests with large margins without increasing the spoiled ballot rate noticeably.

Voters in a hurry, et al.

- Mallory can monitor how long it takes voters to make their selections, whether they change selections, how long they review the summary screen, etc.

Voters in a hurry, et al.

- Mallory can monitor how long it takes voters to make their selections, whether they change selections, how long they review the summary screen, etc.
- A voter who spends little time reviewing selections onscreen also might be unlikely to review the printout carefully.

Voters in a hurry, et al.

- Mallory can monitor how long it takes voters to make their selections, whether they change selections, how long they review the summary screen, etc.
- A voter who spends little time reviewing selections onscreen also might be unlikely to review the printout carefully.
- If a voter takes a very long time to mark a ballot or changes selections repeatedly, might be a sign that the voter finds voting difficult or confusing; such voters might also be unlikely to notice errors in the printout.

FUD attacks on passive testing

- passive testing using the spoiled ballot rate does not produce direct evidence of malfeasance or malfunction
- does not identify which ballots and which contests, if any, have errors
- does not provide any evidence about whether the errors, if any, changed outcomes
- opens the door to a simple, legal way to undermine elections: ask voters to spoil ballots.

Oracle bounds: “shoulder surfing”

- suppose Pat could ask an oracle whether a particular BMD printout had an error (equivalently, suppose Pat can watch over the shoulder of selected voters as they use the BMD)

Oracle bounds: "shoulder surfing"

- suppose Pat could ask an oracle whether a particular BMD printout had an error (equivalently, suppose Pat can watch over the shoulder of selected voters as they use the BMD)
- contest w 2980 voters (2018 median jurisdiction turnout). Mallory alters 15 transactions. Could chance contest outcome by 1% or more.

Oracle bounds: "shoulder surfing"

- suppose Pat could ask an oracle whether a particular BMD printout had an error (equivalently, suppose Pat can watch over the shoulder of selected voters as they use the BMD)
- contest w 2980 voters (2018 median jurisdiction turnout). Mallory alters 15 transactions. Could chance contest outcome by 1% or more.
- Pat would need to spy on $n = 540$ voters, about 18%. Involves testing each BMD several times per hour.

Oracle bounds: "shoulder surfing"

- suppose Pat could ask an oracle whether a particular BMD printout had an error (equivalently, suppose Pat can watch over the shoulder of selected voters as they use the BMD)
- contest w 2980 voters (2018 median jurisdiction turnout). Mallory alters 15 transactions. Could chance contest outcome by 1% or more.
- Pat would need to spy on $n = 540$ voters, about 18%. Involves testing each BMD several times per hour.
- for once-an-hour testing per machine to give 95% chance of catching problem, need >6,580 voters in the contest, almost triple the median turnout in jurisdictions across the U.S., and roughly 20 times the median number of active voters in incorporated areas in the U.S.

Modeling voter behavior

- Pat can't really shoulder surf: needs to model voter behavior

Modeling voter behavior

- Pat can't really shoulder surf: needs to model voter behavior
- requires complete monitoring of surprisingly many voters

Minimax lower bounds

Pat draws an IID training sample of n transactions from P

Confidence	Test Limit	Altered Votes	Bound (millions)
99%	Inf	0.5%	3.73
99%	Inf	1%	3.46
99%	Inf	3%	2.61
99%	Inf	5%	2.04
95%	Inf	0.5%	1.65
95%	Inf	1%	1.57
95%	Inf	3%	1.29
95%	Inf	5%	1.08

Confidence	Test Limit	Altered Votes	Bound (millions)
99%	2000	0.5%	3.87
99%	2000	1%	3.58
99%	2000	3%	2.69
99%	2000	5%	2.09
95%	2000	0.5%	1.67
95%	2000	1%	1.59
95%	2000	3%	1.31
95%	2000	5%	1.10

Complications and frustrations

- the only remedy is a new election

Complications and frustrations

- the only remedy is a new election
- margins are not known when testing happens

Complications and frustrations

- the only remedy is a new election
- margins are not known when testing happens
- tests have uncertainty {#sec:uncertain}

Complications and frustrations

- the only remedy is a new election
- margins are not known when testing happens
- tests have uncertainty {#sec:uncertain}
- requires new systems, extra hardware, additional staff, training w

Complications and frustrations

- the only remedy is a new election
- margins are not known when testing happens
- tests have uncertainty {#sec:uncertain}
- requires new systems, extra hardware, additional staff, training w
- BMDs will still pose special risks of disenfranchising some voters