BRAVO: Ballot-polling Risk-limiting Audits to Verify Outcomes

Mark Lindeman, Philip B. Stark, and Vincent S. Yates

EVT/WOTE 2012
Bellevue, WA
6–7 August 2012

Evidence-Based Elections: Compliance Audits + Materiality Audits

Effective compliance audit

Determine whether the audit trail is trustworthy enough to determine who won.

Was the system, as deployed—including curation of the audit trail—strongly software independent?

If not, do not declare an outcome (nb: danger of DOS attacks).

Effective materiality audit

If the outcome is wrong, correct it (with high probability).

Requires intact audit trail–need to pass compliance audit first.
Might require counting the entire audit trail by hand.

Risk-Limiting Audits for Materiality

- Historically, much debate over how large a sample to start with. Detection paradigm.
- If we want audits to correct wrong outcomes, crucial question is when to *stop* auditing.
- Answer: If there's compelling evidence that outcome is right, stop; else, keep auditing.

What is Compelling Quantitative Evidence?

- What is the biggest chance that—if the outcome is wrong—the audit would observe what it did observe (votes, errors, . . .)?
- If chance is small, implausible that the outcome is wrong; else, keep auditing.
- Eventually, either have strong evidence that the outcome is right, or the whole contest has been counted by hand and correct outcome is known.
- Can guarantee a large probability of correcting the outcome if the outcome is wrong, no matter why it's wrong (if the audit trail is "good enough")
- C.f., Rivest & Shen: Bayesian audits

What is "Risk" in a Risk-Limiting Audit?

The risk is the *largest possible chance* that a wrong outcome will not be corrected by the audit.

The risk-calculation assumes the outcome is wrong in the way that's hardest to detect, *as if a smart, malicious opponent were trying to commit fraud and not get caught*.

Chance of not correcting a wrong outcome is typically much smaller than the risk, e.g., if machine malfunction, configuration error, or voter error is at fault.

The risk is *not* the chance that the outcome, after auditing, is wrong.

For instance, if 99% of outcomes are right in the first place and we audit with a risk limit of 10%, after auditing, on average more than 99.9% of outcomes will be correct:
On average, audit corrects at least 90% of the 1% that are wrong.

Role of Statistics

Limiting the risk is easy

No statistics needed: just count all the ballots by hand.

Statistics lets you do less counting when the outcome is right, but still ensure a big chance of a full hand count when outcome is wrong.

Risk-Limiting Audits

- 16 pilot audits in CA, CO, and OH; another 14 planned (OC next, 8/20)
- EAC funding for pilots in CA and CO and Cuyahoga County, OH
- CO has law; CA has pilot law
- simple measures
- measures requiring super-majority
- multi-candidate contests
- vote-for-*n* contests,
- multiple contests audited simultaneously with one sample
- contest sizes: 200 ballots to 121,000 ballots
- counting burden: 16 ballots to 7,000 ballots
- cost per audited ballot: nil to about $0.55.

Ballot-polling audits and Comparison Audits

- Comparison audit:
  1. LEO “commits” to vote data at some level of aggregation.
  2. Audit checks that the committed data produces the same results as claimed. Should be perfect.
  3. Audit samples and checks the committed data until there is strong evidence that the data are accurate enough to produce the right election outcome (or until the true outcome is known).
- Ballot polling audit: Sample/examine ballots until there is strong evidence that looking at the rest would confirm the outcome (or until the true outcome is known).

Tradeoffs

- Comparison audit
  - Heavy demands on voting system for reporting and export
  - Requires LEO to commit to auditable subtotals
  - Requires ability to retrieve ballots that correspond to CVRs or subtotals
  - May compromise voter privacy (small-batch or ballot-level reporting) & enable coercion through pattern voting
  - Most efficient (ballot-level) may require re-scanning all ballots
  - Checks tabulation (but not for *transitive audits* [Calandrino, Halderman, & Felten] unless subtotals are cross-checked)
  - Ballot-level comparison audits require least hand counting
- Ballot polling audit
  - Requires more counting than *ballot-level* comparison audit
  - Does not check tabulation: Outcome could be right b/c errors cancel
  - Virtually no set-up costs
  - Requires nothing of voting system
  - Generally, need a ballot manifest to draw sample
  - Preserves voter anonymity except possibly for sampled ballots
  - Counting burden comparable to precinct-based comparison audit, unless margin is very small

Counting errors versus counting votes

Johnson (2004): statistical recount versus statistical error count. Like two-sample t-test versus paired t-test.

If constrained to examine batches of a given size, much more efficient statistically (in counting effort) to count errors in those batches than to count votes in those batches.

But if:

- you can only examine precinct-level batches for error
- exporting precinct-level data is hard/complex/time-consuming
- you can examine individual ballots to count votes

then counting votes can be much more efficient overall.

Getting CVRs for Individual Ballots is Hard!

- Federally certified voting systems do not provide CVRs.
- Even getting precinct-level data from today's voting systems into a usable form can take hours of hand editing . . . and then the batch size is too large for efficient audits.
- Today's talk by Kai Wang—the Wang/Wagner et al. software is great, but need LEOs to re-scan ballots, need to program ballot definitions, etc.
  Serious obstacles to ballot-level comparison audits.
- Need ballot manifests for any kind of risk-limiting audit—comparison or ballot-polling.

Ballot-Polling Audit: Intuition

- Like opinion poll or exit poll, but sample until observed winner's percentage (i.e., sample percentage), discounted by "margin of error," is above 50% (for 2-candidate contest).
- If winner's *true* percentage of valid votes is more than 50%, she won.
- If the true margin is in fact small, confirming outcome might require looking at a lot of ballots; if it's big, don't expect to need to see many randomly selected ballots to have strong evidence that the winner got more than 50%.
- E.g., chance the first 4 ballots selected all would show votes for the reported winner if the reported winner didn't get more than 50% of the vote is 6.25% (less than 10%).
- If the true margin is in fact *negative* (i.e., if the reported winner really lost), very unlikely that sample percentage, discounted by "margin of error," will be over 50%.

Wald ballot-polling audit, 2 candidates, risk limit α

1. Pick D, maximum draws before full hand count. Set $T = 1$, $d = 0$. s is winner's share of the valid votes according to the vote tabulation system.
2. Select a ballot at random from ballots cast in the contest. $d \leftarrow d + 1$.
3. If the ballot is an undervote, overvote, or an invalid ballot, return to step 2.
4. If the ballot shows a valid vote for the reported winner, multiply T by

$$2s.$$

5. If the ballot shows a valid vote for anyone else, multiply T by

$$2(1 - s).$$

6. If $T > 1/\alpha$, stop the audit: Reported outcome stands.
   Else if $d < D$, return to step 2.
7. Perform full hand count; hand-count results trump reported results.

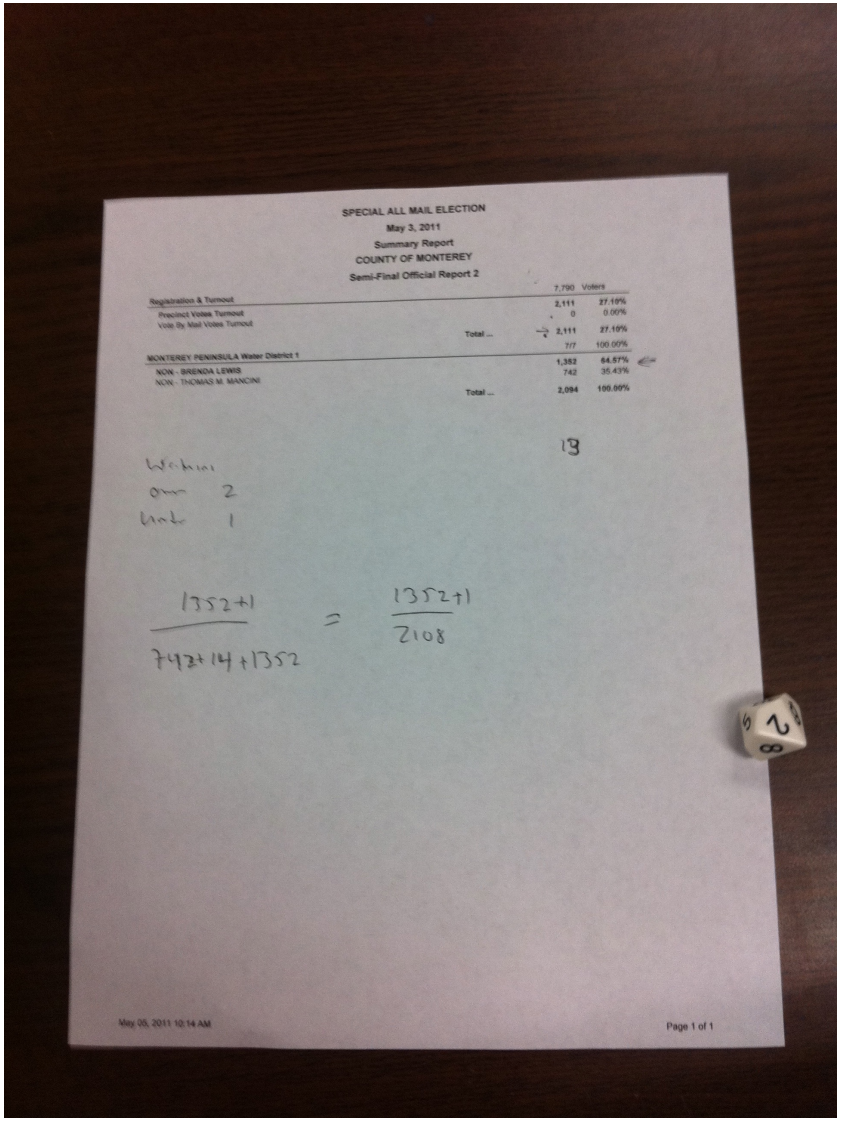
Theorem: limits risk to α.

Ballot-polling audit Monterey Peninsula Water District 1

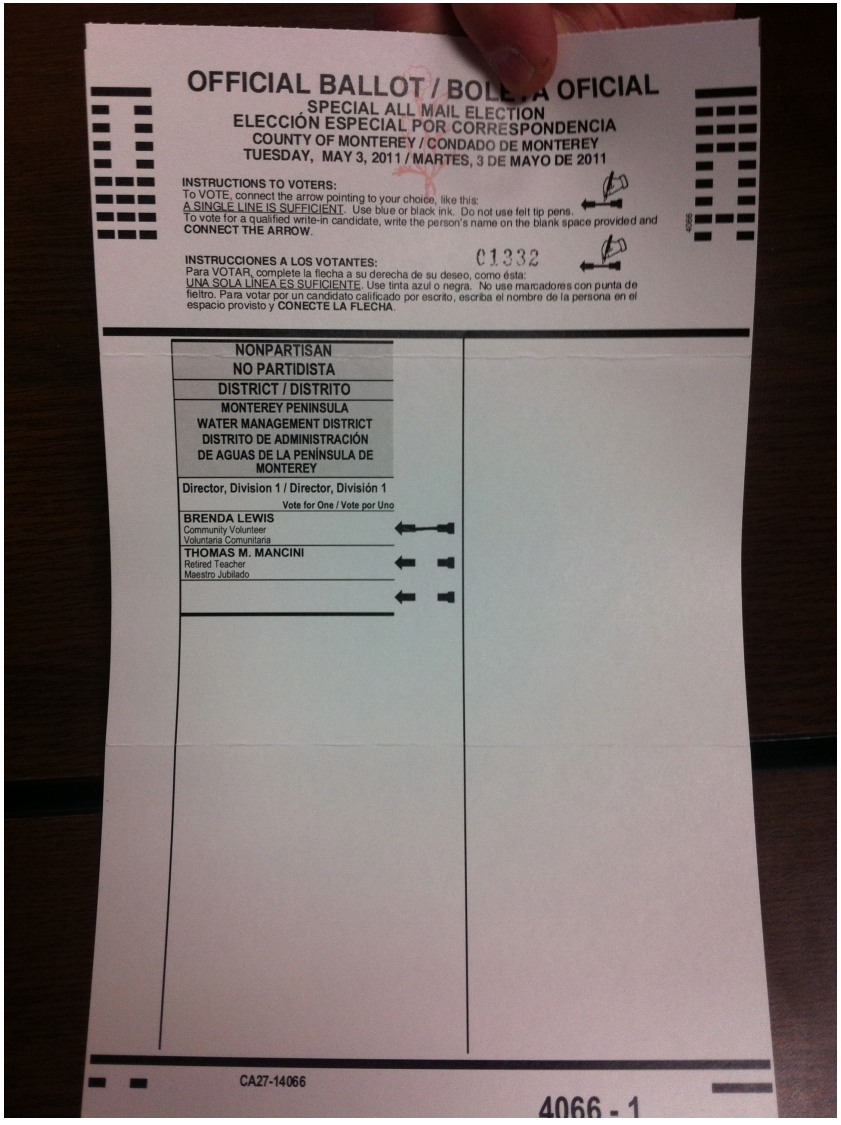
- Conducted in Monterey County in May, 2011, before certification
- 10% risk limit
- Expected number of ballots to examine: 58
- Actual: 92 draws (89 distinct ballots)
- Monterey County staff Bates' stamped every ballot
- Thanks to RoV Linda Tulett & staff!

Monterey County 2011

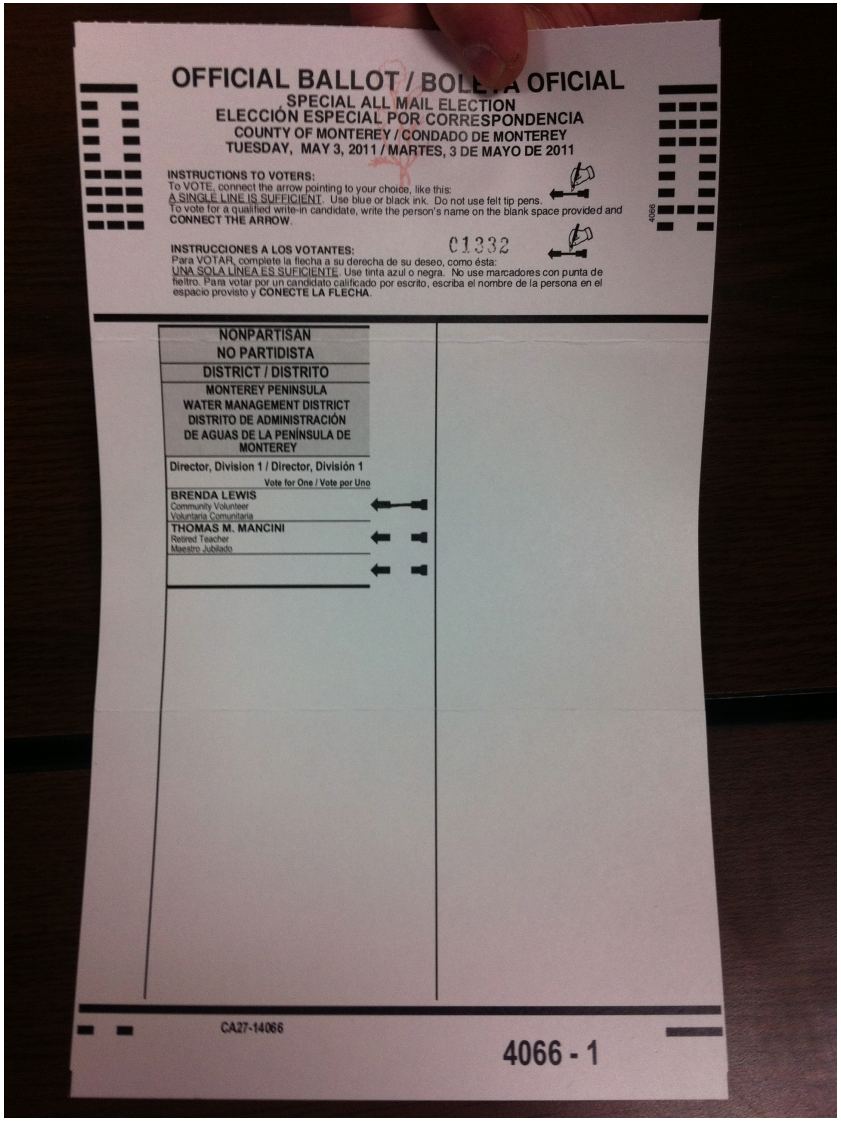
Monterey County 2011



Monterey County 2011



Monterey County 2011



Monterey County 2011

2008 Presidential Contest in CA

Expected sample size to confirm Obama won
Vote share 61.1%:

- $\approx$ 100 ballots from whole state
- $\approx$ 25 from LA County
- $\approx$ 75 total from largest 12 counties (including LA)
- $\approx$ 1 total from the smallest 14 counties.

If Obama's share had been 52%:

- $\approx$ 2,900 from whole state ($\approx$ 0.02% of ballots)
- $\approx$ 725 from LA county
- $\approx$ 2175 total from largest 12 counties (including LA)
- $\approx$ 29 total from smallest 14 counties

Expected Workload: Two Candidates

Winner's True Share	Quantiles					Mean
	25^{th}	50^{th}	75^{th}	90^{th}	99^{th}	
70%	12	22	38	60	131	30
65%	23	38	66	108	236	53
60%	49	84	149	244	538	119
58%	77	131	231	381	840	184
55%	193	332	587	974	2,157	469
54%	301	518	916	1,520	3,366	730
53%	531	914	1,619	2,700	5,980	1,294
52%	1,188	2,051	3637	6,053	13,455	2,900
51%	4,725	8,157	14,486	24,149	53,640	11,556
50.5%	18,839	32,547	57,838	96,411	214,491	46,126

Means and percentiles of #ballots with valid votes to inspect for 10% risk limit. Estimated using 10^7 replications.

Ballot-Polling Audit–k-winner contest: Intuition

Again, like opinion poll or exit poll, but sample until *for every (winner, loser) pair*, observed winner's fraction of votes on ballots that report a vote for either or both, discounted by "margin of error," is above 50%.

That is, sample until there's strong statistical evidence that every reported winner actually got more votes than every reported loser.

If any (winner, loser) margin is in fact small, might require looking at a lot of ballots.

General BRAVO: C-candidate, k-winner contest

Test that every winner $w \in \mathcal{W}$ beat every loser $\ell \in \mathcal{L}$.
$k(C-k)$ null hypotheses: loser ℓ beat winner w.

Test w/ same sample, but one test statistic per pair: $\{T_{w\ell}\}$.

Define $s_{w\ell} \equiv s_w/(s_w + s_\ell)$, fraction of votes w was reported to have received among ballots reported to show a vote for w or ℓ or both.

Can be calculated from standard reported election results.

Define $\pi_{w\ell}$ to be actual fraction of votes w received among ballots that show a vote for *exactly one* of $\{w, \ell\}$.

Assertion and Sufficient Condition

$\forall w \in \mathcal{W}, \ell \in \mathcal{L}$:

- If w reportedly beat ℓ, $s_{w\ell} > 50\%$.
- If w actually beat ℓ, $\pi_{w\ell} > 50\%$.

BRAVO for C-candidate k-winner contest

1. Pick D. Set $d = 0$ and set $T_{w\ell} = 1$ for all $w \in \mathcal{W}$ and $\ell \in \mathcal{L}$.
2. Draw a ballot uniformly at random w/ replacement from those cast in contest. $d \leftarrow d + 1$.
3. If the ballot shows a valid vote for a reported winner w, then for each ℓ in $\mathcal{L}$ that did not receive a valid vote on that ballot, multiply $T_{w\ell}$ by $2s_{w\ell}$. Repeat for all such w.
4. If ballot shows a valid vote for a reported loser ℓ, then for each w in $\mathcal{W}$ that did not receive a valid vote on that ballot multiply $T_{w\ell}$ by $2(1 - s_{w\ell})$. Repeat for all such ℓ.
5. For all (w, ℓ) with $T_{w\ell} \geq 1/\alpha$, conclude that w beat ℓ.
   Don't update those $T_{w\ell}$ further.
6. If have concluded that all $w \in \mathcal{W}$ beat all $\ell \in \mathcal{L}$, stop: Reported results stand.
   Else if $d < D$, return to step 2.
7. Perform full hand count; results replace the reported results.

Theorem: Limits risk to at most α.

Steampunk audit

Equipment needed: dice, pencil and paper (or a sliderule).

Calculations very transparent (even if underlying theorems are hard).

Process very observable: What votes does this ballot show?

Multiplicity in pairwise testing for k-winner contest

Stopping short of a full hand count is an error only if at least one of the null hypotheses is in fact true.

BRAVO stops short of full hand count only if all $k(C - k)$ null hypotheses are rejected.

Consider the set of null hypotheses that are true. Chance of erroneously rejecting *all* of those is at most the smallest chance of erroneously rejecting any individually.

Hence, by testing every (winner, loser) pair individually at level α, the chance of stopping short of a full hand count if any of the $C - k$ apparent losers actually won is at most α.

Moreover, works simultaneously for any number of contests, using the same sample.

Grouping losers

Could combine subsets of winners or of losers to reduce the number of tests.

E.g., winner has 60%, losers have 25% and 15%. Combine losers into a single fictitious losing candidate with 40%.

Theorem: grouping does not reduce expected sample size.

Workload at 10% Risk Limit

255 state presidential contests between 1992 and 2008
median statewide expected sample size to confirm the plurality winner in each state using BRAVO is

307 ballots

(On the assumption that the outcomes were right.)

Selecting ballots at random

For transparency, want initial mechanical source of randomness (Cordero, Wagner, & Dill).

Dice courtesy of Ron Rivest.

Use as Seed in Good PRNG

SHA-256 of seed catenated with sample number (Rivest)

Random sampling

Pseudo-Random Sample of Ballots

Seed: 73567556725160627585

Number of ballots: 7116

Current sample number: 623

Draw this many ballots: 623 draw sample reset

Ballots selected: ☑ show sequence numbers ☐ show hash values

```
sequence_number, ballot
1,2086
2,2462
3,3320
4,4719
5,4813
6,3838
7,2655
8,2747
9,3059
```

Ballots selected, sorted:

```
19,34,37,38,51,90,96,96,99,101,109,114,150,156,163,175,187,187,195,197,198,244,280,281,301,316,
372,395,403,404,407,417,429,444,450,451,471,477,480,481,482,491,514,542,545,550,554,577,585,58
5,596,597,613,614,615,629,645,647,657,685,692,692,694,739,750,763,768,792,795,798,819,832,841,
842,857,862,871,874,876,884,901,906,923,923,934,937,937,958,963,973,978,1018,1049,1050,1071,1
081,1097,1105,1125,1126,1130,1165,1205,1210,1218,1219,1224,1226,1284,1288,1291,1318,1327,13
57,1370,1372,1388,1406,1422,1425,1432,1433,1434,1446,1447,1457,1484,1494,1496,1507,1512,152
3,1524,1540,1572,1574,1575,1576,1611,1614,1626,1634,1638,1642,1644,1665,1677,1685,1718,1735,
1761,1764,1774,1788,1791,1793,1816,1827,1851,1855,1893,1921,1978,1989,2010,2017,2034,2056,2
056,2058,2062,2069,2083,2086,2100,2112,2152,2189,2192,2206,2208,2210,2213,2224,2249,2266,22
91,2295,2302,2331,2332,2390,2391,2395,2398,2401,2422,2436,2462,2463,2474,2495,2513,2514,252
```

Ballots selected, sorted, duplicates removed:

```
19,34,37,38,51,90,96,99,101,109,114,150,156,163,175,187,195,197,198,244,280,281,301,316,372,395
,403,404,407,417,429,444,450,451,471,477,480,481,482,491,514,542,545,550,554,577,585,596,597,6
13,614,615,629,645,647,657,685,692,694,739,750,763,768,792,795,798,819,832,841,842,857,862,871
,874,876,884,901,906,923,934,937,958,963,973,978,1018,1049,1050,1071,1081,1097,1105,1125,1126
,1130,1165,1205,1210,1218,1219,1224,1226,1284,1288,1291,1318,1327,1357,1370,1372,1388,1406,1
422,1425,1432,1433,1434,1446,1447,1457,1484,1494,1496,1507,1512,1523,1524,1540,1572,1574,15
75,1576,1611,1614,1626,1634,1638,1642,1644,1665,1677,1685,1718,1735,1761,1764,1774,1788,179
1,1793,1816,1827,1851,1855,1893,1921,1978,1989,2010,2017,2034,2056,2058,2062,2069,2083,2086,
2100,2112,2152,2189,2192,2206,2208,2210,2213,2224,2249,2266,2291,2295,2302,2331,2332,2390,2
391,2395,2398,2401,2422,2436,2462,2463,2474,2495,2513,2514,2520,2549,2556,2558,2563,2578,25
```

Ballot Manifest

Find ballots using a ballot manifest

Ballot look-up tool

Ballot manifest: Each line must have a batch label, a comma, and one of the following:
(i) the number of ballots in the batch
(ii) a range specified with a colon (e.g., 131:302), or
(iii) a list of ballot identifiers within parentheses, separated by spaces (e.g., (996 998 1000)).
Each line should have exactly one comma.

```
001_211161_01,23
002_211162_02,9
003_211561_03,32
004_211561_03,50
005_211561_03,50
006_211562_04,14
007_211562_04,50
008_211562_04,50
009_211562_04,50
010_211563_05,12
011_211751_06,27
012_211761_07,2
013_211761_07,50
014_211761_07,50
015_211761_07,50
016_211761_07,50
017_211771_08,2
018_221161_09,16
019_221161_09,50
020_221161_09,50
021_221161_09,50
022_221162_10,30
023_221162_10,50
024_221162_10,50
025_221162_11,50
```

Ballots to look up (separated by commas):

```
19,34,37,38,51,90,96,96,99,101,109,114,150,156,163,175,187,187,195,197,198,244,280,281,301,316,
372,395,403,404,407,417,429,444,450,451,471,477,480,481,482,491,514,542,545,550,554,577,585,58
5,596,597,613,614,615,629,645,647,657,685,692,692,694,739,750,763,768,792,795,798,819,832,841,
842,857,862,871,874,876,884,901,906,923,923,934,937,937,958,963,973,978,1018,1049,1050,1071,1
081,1097,1105,1125,1126,1130,1165,1205,1210,1218,1219,1224,1226,1284,1288,1291,1318,1327,13
57,1370,1372,1388,1406,1422,1425,1432,1433,1434,1446,1447,1457,1484,1494,1496,1507,1512,152
3,1524,1540,1572,1574,1575,1576,1611,1614,1626,1634,1638,1642,1644,1665,1677,1685,1718,1735,
1761,1764,1774,1788,1791,1793,1816,1827,1851,1855,1893,1921,1978,1989,2010,2017,2034,2056,2
056,2058,2062,2069,2083,2086,2100,2112,2152,2189,2192,2206,2208,2210,2213,2224,2249,2266,22
91,2295,2302,2331,2332,2390,2391,2395,2398,2401,2422,2436,2462,2463,2474,2495,2513,2514,252
```

Look-up

look up ballots

Sorted lookup table:

sorted_number, ballot, batch_label, which_ballot_in_batch
1, 19, 001_211161_01, 19
2, 34, 003_211561_03, 2
3, 37, 003_211561_03, 5
4, 38, 003_211561_03, 6
5, 51, 003_211561_03, 19
6, 90, 004_211561_03, 26
7, 96, 004_211561_03, 32
8, 96, 004_211561_03, 32
9, 99, 004_211561_03, 35
10, 101, 004_211561_03, 37
11, 109, 004_211561_03, 45
12, 114, 004_211561_03, 50
13, 150, 005_211561_03, 36
14, 156, 005_211561_03, 42
15, 163, 005_211561_03, 49
16, 175, 006_211562_04, 11
17, 187, 007_211562_04, 9
18, 187, 007_211562_04, 9
19, 195, 007_211562_04, 17
20, 197, 007_211562_04, 19
21, 198, 007_211562_04, 20
22, 244, 008_211562_04, 16
23, 280, 009_211562_04, 2
24, 281, 009_211562_04, 3

Gotchya!

Better ballot accounting

Ballot manifests are *not* a solved problem.

It's easy to deal with errors in ballot manifest if there's an upper bound on the number of ballots in each container (Bañuelos & Stark).

But sometimes there isn't a good upper bound—esp. with multipage ballots.

GOTA: Get out the Audit!

Ballot-polling audits are possible for the November 2012 presidential election in any jurisdiction that has VVPRs—and has knows how many and where they are.

Workload *not* large in most states; preparations minimal.
Equipment needed: dice, pencil, and paper.
(Alternatively, dice and simple web-based tools.)

Compliance audit needs attention—ensure audit trail adequately accurate.
Coordination across jurisdictions needs attention—logistics and transparency.

Verified Voting is working to get ballot-polling audits in several states for November 2012 presidential election.

Let's Get out the Audit!