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

Mark Lindeman
Philip B. Stark[1]
Vincent S. Yates[1]
[1]*Department of Statistics*
University of California, Berkeley

Abstract

Risk-limiting post-election audits guarantee a high probability of correcting incorrect electoral outcomes, regardless of why the outcomes are incorrect. Two types of risk-limiting post-election vote tabulation audits are *comparison audits* and *ballot-polling audits*. Comparison audits check some of the subtotals reported by the vote tabulation system, by hand-counting votes on the corresponding ballots. Ballot-polling audits select ballots at random and interpret those ballots by hand until there is strong evidence that the outcome is right, or until all the votes have been counted by hand: They directly assess whether the outcome is correct, rather than assessing whether the tabulation was accurate. Comparison audits have advantages, but make large demands on the vote tabulation system. Ballot-polling audits make no such demands. For small margins, they can require large samples, but the total burden may still be modest for large contests, such as county-wide or state-wide races. This paper describes BRAVO, a flexible protocol for risk-limiting ballot-polling audits. Among 255 state presidential contests between 1992 and 2008, the median expected sample size to confirm the plurality winner in each state using BRAVO was 307 ballots (per state). Ballot-polling audits can improve election integrity immediately at nominal incremental cost to election administration.

1 Introduction and background

Voting systems in use in the U.S. are known to be vulnerable to misinterpretation of voter intent, mechanical failure, misconfiguration, and deliberate subversion [Secretary of State, 2007; McDaniel et al., 2007]. The vote totals they produce should not be assumed to be accurate, nor even sufficiently accurate to produce correct outcomes. The possibility of gross error is not merely theoretical: In March 2012, a routine post-election audit in Palm Beach County, Florida led to a recount that found the wrong winners had been declared in two city council contests [Sorentrue et al., 2012]. Wrong outcomes in U.S. elections are thought to be rare, but the true incidence is unknown.

Risk-limiting post-election audits have been widely endorsed as a means to check whether voting systems find the correct outcomes [ElectionAudits.org, 2008]. A risk-limiting audit checks some voted ballots (or voter-verifiable records) in search of strong evidence that the reported election outcome was correct. If the reported outcome is *incorrect*, a risk-limiting audit has a large, pre-specified minimum chance of leading to a full hand count, which reveals the correct outcome. The *risk limit* is the maximum chance that an audit of an election with an incorrect outcome will *not* lead to a full hand count. A burgeoning literature (e.g., Checkoway et al. [2010]; Hall et al. [2009]; Stark [2008a,b, 2009a,c,b, 2010]; Benaloh et al. [2011]; Lindeman and Stark [2012]) explores methods and concomitants for risk-limiting audits.

Two types of risk-limiting audits have been proposed: *comparison audits* and *ballot-polling audits* (Lindeman and Stark [2012]; Johnson [2004] makes an analogous distinction). Comparison audits check outcomes by comparing hand counts[1] to voting system counts for clusters of ballots.[2] In a *ballot-level* comparison audit, each cluster is a single ballot. In a *batch-level* comparison audit, each cluster comprises multiple ballots: for instance, all the ballots cast in a particular precinct. In either case, a risk-limiting comparison audit examines randomly selected clusters to assess whether the voting system subtotals are sufficiently accurate to confirm the reported out-

[1] We construe "counting" the votes in a cluster of ballots to have two parts: (i) Interpreting individual ballots or records in the cluster to identify valid votes. (ii) Tallying the valid votes on those ballots. A hand count of a cluster that consists of a single ballot cast in a plurality contest yields a one or a zero for each candidate, a one if the ballot shows a valid vote for the candidate or a zero if not.

[2] *Transitive audits* use counts from a secondary system, as briefly discussed below.

come, at the specified risk limit. If not, a full hand count is conducted.

In contrast, a ballot-polling audit does not use voting system subtotals. Instead, it examines the votes on a random sample of individual ballots.[3] When and if the vote shares in the sample provide sufficiently strong evidence to confirm the reported outcome, the audit stops.

Comparison risk-limiting audits have important benefits, but they can be hard to implement. The most efficient comparison audits—those at the ballot level—require information that current certified vote tabulation systems do not provide: interpretations of individual ballots (cast-vote records or CVRs) that can be associated with the paper ballots they purport to represent. Preparing for a comparison audit requires exporting a complete list of auditable subtotals from the voting machine and the ability to retrieve (and hand count) the ballots corresponding to each subtotal. It sometimes requires translating voting-system reports that were intended for printing into formats that software can read, a labor-intensive, error-prone process. And it requires summing the auditable subtotals to verify that they match the reported contest totals.

To conduct a ballot-level comparison audit of a current certified system generally requires *transitive auditing* (see Calandrino et al. [2007], although they do not call it by that name). A transitive audit uses a secondary system to make a CVR for each ballot in a way that allows the CVR to be matched to the ballot it purports to represent. So far, transitive audits have relied on digital images of all the ballots cast in the contest (produced by the voting system or by rescanning the ballots); creating CVRs from those images using a combination of software and hand labor; and maintaining a mapping between the physical ballots and the CVRs by keeping the ballots in the order in which they were scanned or by marking individual ballots with an identifier. If the outcome according to the secondary system matches the voting system's reported outcome, then a risk-limiting audit of the secondary system outcome is a risk-limiting audit of the official system, at the same risk limit: The reported outcome is incorrect if and only if the secondary system's outcome is incorrect. Hence, if the reported outcome is incorrect, the chance that the audit will lead to a full hand count of the physical ballots is at least 100% minus the risk limit used in the transitive audit. While the transitive audit strategy reaps the statistical efficiency advantages of ballot-level comparison audits, the logistical obstacles may be substantial, especially in large jurisdictions.

Ballot-polling audits, in contrast, require virtually nothing from the voting system and no extra preparation beyond the sorts of ballot accounting that local election officials generally do. They are an immediate option in any jurisdiction with an auditable paper trail. Moreover, for most margins of victory, they typically require examining substantially fewer ballots than batch-level comparison audits to attain a particular risk limit. Thus, we propose ballot-polling audits as a practical way to conduct risk-limiting audits of selected contests—such as presidential elections—immediately.

When the reported count is accurate or nearly accurate, the amount of auditing to be done in a risk-limiting ballot-polling audit is less predictable than the amount to be done in a risk-limiting comparison audit of the same contest. Therefore it is advantageous to be able to expand the audit flexibly depending on early results, using *sequential sampling*. The best-known sampling methods use a single sample and cannot easily be extended to risk-limiting sequential audits. The method we describe here is based on a robust sequential sampling method that can draw one or more ballots at a time, as the auditors may desire.

2 Previous work

Johnson [2004] presents the first ballot-polling election audit of which we are aware. Johnson calls his ballot-polling method a "statistical recount."[4] To conduct a statistical recount, an initial random sample of n ballots is drawn without replacement. If the reported winner's margin in the sample is larger than a "critical margin" (essentially half the width of a confidence interval at some predetermined confidence level), the audit stops and the election is "validated." If the sample margin is negative and larger in absolute value than the critical margin—giving evidence that the loser actually won—the audit stops and the election is "invalidated." If the absolute value of the sample margin is smaller than the critical margin, the audit expands the random sample in stages. At each stage, a random sample of n ballots is drawn without replacement from the as-yet-unexamined ballots. The cumulative sample margin is compared with a critical margin (updated to reflect the cumulative sample size), until the election outcome has been validated or invalidated, or all ballots have been examined. Johnson asserts that the method can be used to validate election outcomes with a specified level of "statistical confidence." However, the distribution of the maximum of a collection of (normalized) hypergeometric random variables with nested samples is not (normalized) hypergeometric, which the method assumes implicitly. This could make the chance that the method mistakenly validates an

[3] Batch-level ballot-polling audits are possible, but less efficient than all of the approaches discussed here. They do not seem to have any advantage over ballot-level ballot-polling audits.

[4] Johnson [2004] also discusses a ballot-level comparison audit, which he calls a "statistical error count."

incorrect outcome much larger than 1 minus the confidence level. Hence, a "statistical recount" is not risk-limiting.

Simon and O'Dell [2006] propose to verify federal election outcomes through "Universal Precinct-based Sampling" (UPS). UPS entails drawing and hand-counting a 10% sample of the ballots cast in each precinct, and each other "pile" of ballots—preferably on election night. Simon and O'Dell assert that these samples (treated as simple random samples) combined provide vote share estimates "within 1% of an accurate vote count" with 99% statistical confidence in competitive House contests—and even more accurate estimates in larger contests. Discrepancies of more than 1% in a reported winner's vote share would trigger further scrutiny, i.e., further sampling or perhaps a full hand count. UPS is not risk-limiting, for a variety of reasons.[5] UPS emphasizes immediacy, whereas BRAVO emphasizes rigor, limiting risk, and efficiency: BRAVO requires dramatically smaller samples in many large contests, as we demonstrate below.

Lindeman and Stark [2012] sketch a risk-limiting ballot-polling audit method similar to BRAVO. BRAVO improves on that method in a number of ways: (1) Auditors are never required to escalate to a full hand count based on a test statistic. (2) That method applies only to vote-for-one contests with a majority winner, while BRAVO works with vote-for-one or vote-for-*k* plurality, majority, or supermajority contests. (3) Because BRAVO uses pairwise comparisons instead of "pooling" votes for reported losing candidates, it can be substantially more efficient.

There is by now a large literature on sequential testing, but to our knowledge, developments since Wald [1945] are not helpful for risk-limiting ballot-polling audits. For instance, there are sequential procedures for identifying the "best" multinomial category by sampling until one category has occurred substantially more frequently than any other [Gibbons et al., 1977; Ng and Panchapakesan, 2007; Ramey and Alam, 1979]. These are not suited to risk-limiting audits: They give a final *P*-value associated with the "best" category, but provide no mechanism to assure that this *P*-value is below the desired risk limit α.

Rivest and Shen [2012] propose a Bayesian approach to election auditing that can be used with ballot polling. It is not clear whether this approach is risk-limiting, nor how the risk limit might depend on the prior probability distribution.

3 Overview of BRAVO

We assume that the election generated an indelible, voter-verifiable audit trail, which might consist of a combination of voter-marked paper ballots and voter-verifiable paper records (VVPRs). We refer to such records as "ballots," even though they might not be paper ballots. We also assume that a *compliance audit* [Benaloh et al., 2011; Lindeman and Stark, 2012; Stark and Wagner, 2012] has provided convincing evidence that this audit trail is adequately accurate to reflect who actually won; otherwise, a full hand count—the recourse of risk-limiting audits when the outcome is in doubt—would not necessarily reveal the actual winner.

The ballot-polling risk-limiting audit we describe here, in its simplest form, can be described as a loop:

1. Randomly select a ballot and examine it for a valid vote. (Ballots can be selected more than once; they are counted as many times as they are selected.)

2. If the ballots selected so far provide sufficiently strong evidence that the reported outcome is right, stop the audit and accept the reported outcome.

3. If, after examining a large number of ballots, the audit has not given strong evidence that the reported outcome is right—or if it provides strong evidence that the reported outcome is wrong—conduct a full hand count of all the ballots to determine the correct outcome.

4. If neither condition 2 nor 3 holds, return to step 1.

For the audit to be risk-limiting, condition 2—the criterion of "sufficiently strong evidence"—must be defined correctly before the audit. Condition 3 can be less sharply defined; what matters is that the audit ends either by meeting condition 2 or by conducting a full hand count.

The arithmetic involved in BRAVO is simple enough to do on a four-function calculator. The ballotPollAuditTools web page (`http://statistics.berkeley.edu/~stark/Java/Html/ballotPollTools.htm`), or custom tools, can be used to help with the sampling logistics and calculations involving multiple candidates.

A sequential audit of one ballot at a time can reduce the expected number of ballots that must be inspected

[5] (1) As described, the chance UPS leads to a full hand count when the outcome is incorrect could be zero, so the method is not risk-limiting. (2) The fixed 10% sample size does not ensure that the standard error of the sample margin will be small enough to draw a conclusion about the sign of the true margin. UPS could fall far short of its nominal 99% confidence level. (3) UPS uses a stratified sample, so Simon and O'Dell's margin of error calculation is at best approximate. An exact calculation would require knowing how many ballots are in each "pile," and we suspect that the calculation would be intractable, although conservative bounds might be obtained. (4) One sampling method Simon and O'Dell propose—examine every tenth ballot, starting with the *n*th ballot, where *n* is a random integer between 1 and 10 inclusive—further complicates the risk calculation compared with taking independent simple random samples from each "pile."

when the apparent outcome is correct, because the audit can stop as soon as there is strong evidence that the apparent winner(s) actually won. However, it may be more efficient to select and inspect more than one ballot at a time, especially to audit contests that cross jurisdictional boundaries. Accordingly, we suggest methods to conduct the audit in groups or "stages" of ballots so that jurisdictions can audit concurrently.

4 Applicability of BRAVO

We address simple measures, measures that require a super-majority, and plurality contests, including contests such as city council contests and school board contests in which each voter may select more than one candidate. The method we present is not suitable for auditing ranked-choice voting (RCV) or instant-runoff voting (IRV), because whether an RCV or IRV ballot ultimately is counted for a particular candidate typically depends on the other ballots and cannot be determined in isolation.

There are C candidates in the contest; there are $k \geq 1$ reported winners. In a simple measure, $C = 2$ and $k = 1$. In a city council contest, we might have $C = 10$ and $k = 5$. For plurality contests, determining whether the k reported winners among the C candidates in the contest really won amounts to determining whether each of the k reported winners received more votes than all of the $C-k$ reported losers. As elaborated in Stark [2008a], it is natural to take the null hypothesis to be that the outcome is wrong; in this case, that means that at least one of the reported losers received at least as many votes as at least one of the reported winners. To reject that hypothesis is to conclude that all k candidates reported to have won really did win. BRAVO tests this hypothesis by testing the union of the hypotheses that apparent loser ℓ received at least as many votes as apparent winner w, for all pairs (w, ℓ) of apparent winners and apparent losers. Testing this composite hypothesis at significance level α limits the risk to α, if a full hand count is required whenever the test does not (eventually) reject all of the pairwise hypotheses.[6]

For contests that require an outright majority, testing whether the apparent winner truly won amounts to testing the null hypothesis that no more than half of the valid votes are for the reported winning position. To reject that hypothesis is to conclude that the apparent majority position really received a majority (¿ 50%) of the votes. As before, testing at significance level α limits the risk to α, if a full hand count is required whenever the test does not ultimately reject the null hypothesis. The test can be generalized to a supermajority requirement by replacing 50% with the appropriate proportion.

There are many possible approaches to ballot-polling risk-limiting audits. The approach we present here is based on Wald's Sequential Probability Ratio Test (SPRT) [Wald, 1945]. In Wald's method, sampling continues until either the null hypothesis or the alternative hypothesis is accepted. In BRAVO, the null hypothesis that the reported election outcome is wrong is never be accepted based on anything short of a full hand count. For plurality contests with more than two candidates, BRAVO performs several SPRTs in parallel using the same sample, to test whether any loser is in fact a winner. Methods sharper than BRAVO could be devised, although we doubt that the computations would be as simple.

5 Sampling framework and notation conventions

We assume that the contest under audit is a plurality contest with C candidates and $k \geq 1$ winners. The set $\mathscr{W}$ denotes the the apparent winners and the set $\mathscr{L}$ denotes the apparent losers. The set $\mathscr{W}$ contains k candidates. (Generally, a voter is permitted to vote for up to k candidates if the contest has k winners, but that restriction matters only to determine whether a given ballot has valid votes in the contest: A ballot marked for too many candidates would be treated as an overvote rather than as a valid vote for each of those candidates.) Measures that require a majority or supermajority are treated incidentally in section 6. We assume that every candidate in $\mathscr{W}$ was reported to have more votes than every candidate in $\mathscr{L}$; that is, the reported winner with the fewest votes is not apparently tied with the reported loser with the most votes.[7]

The true, unknown proportion of ballots that record votes for candidate c is π_c. We consider ballots that do not show a valid vote in the contest (for instance, ballots with overvotes) to be votes for a fictitious candidate 0; candidate 0 is neither in $\mathscr{W}$ nor in $\mathscr{L}$. Thus, the vector $\boldsymbol{\pi_c} \equiv (\pi_c)_{c=0}^{C}$. The reported proportion of ballots that record votes for candidate c is p_c. Typically, instead of reporting p_c directly, election officials report s_c, the fraction of valid votes cast for candidate c; the denominator is not the total number of ballots but rather the total number of reported valid votes. (In a vote-for-one contest, s_c typically exceeds p_c, because some ballots do not show a vote for any candidate in the contest.) The relationship between them is

$$s_c \equiv \frac{p_c}{\sum_{j=1}^{C} p_j}. \tag{1}$$

[6] As elaborated below, the fact that counting continues until *all* the pairwise hypotheses are rejected results in a very simple test that does not need any statistical adjustment for multiplicity.

[7] Statistical sampling is not well suited to confirm that two candidates received exactly the same number of votes.

To check whether the set $\mathscr{W}$ is really the set of winners, we will draw ballots uniformly at random and interpret the votes on those ballots manually. The probability that a ballot selected at random records a vote for candidate c is π_c. The apparent winners really won if

$$\min_{c \in \mathscr{W}} \pi_c > \max_{c \in \mathscr{L}} \pi_c. \tag{2}$$

Equivalently,

$$\pi_w > \pi_\ell, \forall w \in \mathscr{W}, \ell \in \mathscr{L}. \tag{3}$$

We assume that the value of π_0, the proportion of ballots that do not show a valid vote, is irrelevant to the outcome.[8] Thus, the parameter of interest is $\boldsymbol{\pi} \equiv (\pi_c)_{c=1}^{C}$, the vector of conditional probabilities that a ballot bears a valid vote for candidate c given that it bears a valid vote for some candidate. Henceforth, subscripted values of π (e.g., π_i) refer to values in $\boldsymbol{\pi}$, not in $\boldsymbol{\pi_c}$. (However, $\boldsymbol{\pi_c}$ matters for expected sample sizes, especially in the vote-for-one case.)

6 Special case: Contests with two candidates (and majority contests)

To begin, consider the case of a vote-for-one contest with two candidates: the reported winner, w, and the reported loser, ℓ. (A vote-for-one majority contest in which candidate w is reported to have received more than a majority of valid votes can also be audited as described in this section, by treating votes for all other candidates as a vote for ℓ; however, it is more efficient to keep the losing candidates separate, as proved below.) We want to test the null hypothesis that $\pi_w \leq 1/2$. To incorrectly reject the null hypothesis is to keep the reported outcome when it is wrong, so the significance level α is the risk limit.

We want to avoid unnecessary full hand counts, especially when the original vote totals are correct or very nearly so. By the same token, if the audit produces inconclusive results after extensive sampling—or if it produces strong evidence that the reported outcome in fact is incorrect—then we should proceed to a full hand count. Several mechanisms for triggering a full hand count are possible.[9] Here we assume that the auditors optionally set M, a maximum number of ballots to audit, beyond which they intend to proceed to a full hand count if the audit has not produced sufficiently strong evidence that the reported outcome is correct. The value M is a soft limit: It is acceptable to continue the audit after M has been reached,[10] and it is certainly acceptable to proceed to a full hand count before M is reached.

Given s_w (the reported proportion of valid votes cast for w) and the risk limit α, the following steps define a risk-limiting audit based on Wald's SPRT (Wald [1945]):

1. Set the test statistic $T = 1$, and the cumulative sample size $m = 0$. (Optionally, pick M, the maximum number of ballots to audit before requiring a full hand count.)

2. Draw a ballot at random from the set of ballots that include the contest under audit. (A ballot can be drawn more than once.) If the ballot shows a vote for w, multiply T by $s_w/0.5$. If the ballot shows a vote for ℓ, multiply T by $(1-s_w)/0.5$. (If the ballot does not show a valid vote, T is unchanged, or equivalently, multiplied by 1.) Increment m.

3. If $T \geq 1/\alpha$, stop the audit: The reported outcome is confirmed at risk limit α.

4. If $m = M$, or at the auditors' discretion, stop and perform a full hand count; the outcome according to the hand count replaces the reported outcome.

5. If neither (3) nor (4) holds, go back to step (2).

Because $s_w/0.5 > 1$ and $(1-s_w)/0.5 < 1$, T increases each time the audit finds a vote for w and decreases each time it finds a vote for ℓ. (The critical value $1/\alpha$ follows from equation (3.24) in Wald [1945], with $\beta = 0$.) If the reported vote shares are correct, values of T close to 0 are unlikely: The chance of observing $T \leq p$ is no greater than p.[11] Thus, very small values of T provide a rationale for performing a full hand count before M is reached.

If the m ballots drawn so far show m_w votes for w and m_ℓ votes for ℓ, then

$$T = (2s_w)^{m_w}(2-2s_w)^{m_\ell}. \tag{4}$$

Suppose that the reported vote shares are correct. Table 1 gives estimated means and selected percentiles for the expected number of ballots with valid votes to inspect in order to confirm reported outcomes at risk limit $\alpha = 10\%$, assuming that the reported vote shares are correct and that every audit proceeds until the reported outcome is confirmed.[12] The simulations summarized in Table 1 sequentially draw a ballot at random and multiply

[8] If the voting rules dictate otherwise, the method here can readily be adjusted.

[9] Lindeman and Stark [2012] describes the use of β, which strictly controls the chance of a full hand count on the condition that the winner's vote share is accurate within a certain specified tolerance, such that the reported winner in fact won. Wald's SPRT [Wald, 1945] uses both α and β; in the exposition here, implicitly we have set $\beta = 0$.

[10] Auditors may strongly prefer to continue sampling if the audit is close to reaching the risk limit.

[11] This follows from inequality (3.17) in Wald [1945].

[12] If the reported vote shares are even approximately correct, then the expected value of the multiplier in step (2) is positive, so $p(T \geq 1/\alpha) \rightarrow 1$ as $m \rightarrow \infty$.

the test statistic T by $s_w/0.5$ if the ballot shows a vote for the winner or by $(1-s_w)/0.5$ if it shows a vote for the loser. This is repeated until the test statistic $T \geq 1/\alpha$; the number of draws required is recorded. The entire process was repeated 10^7 times for each margin considered. Table 1 reports the sample mean and empirical percentiles of the 10^7 simulations for each margin considered.

Most of the mean sample sizes reported in the table are modest in the context of a large jurisdiction with hundreds of thousands or millions of ballots cast. The mean sample size varies roughly as the inverse square of the margin, $\pi_w - \pi_\ell$; it increases sharply as π_w approaches 50% from above. Notice that the number of ballots to be audited is unpredictable: A small fraction of audits inspect several times more ballots than average.

Table 1 also reports an estimate of the Average Sample Number (ASN) [Wald, 2004], the expected number of draws required either to accept or to reject the null hypothesis. The entries in the table are computed on the assumption that the reported vote totals are correct. Our estimate is

$$ASN \approx \frac{\ln(1/\alpha) + z_w/2}{p_w z_w + p_\ell z_\ell}, \tag{5}$$

where $z_w \equiv \ln(2s_w)$ and $z_\ell \equiv \ln(2-2s_w)$.[13]

Let $s_\ell \equiv 1 - s_w$ and $x = s_w - s_\ell$, the reported margin as a proportion of valid votes. Suppose $p_0 = 0$. Then the denominator of expression (5) becomes

$$\begin{aligned} &\frac{(1+x)\ln(1+x) + (1-x)\ln(1-x)}{2} \\ &= x^2/2 + x^4/12 + x^6/30 + \cdots. \end{aligned} \tag{6}$$

The first term dominates when x is small, so

$$ASN \approx 2\ln(1/\alpha)/x^2. \tag{7}$$

That is, the ASN is roughly inversely proportional to the margin squared.

When there is a positive probability p_0 that a draw will give an invalid ballot or a ballot with an overvote, that "thins" the rate of drawing informative ballots, increasing ASN by the factor $1/(1-p_0) > 1$. For instance, if 10% of ballots contain invalid votes, then the ASN increases by $1/(1-0.1) - 1 \approx 11.1\%$.

When $\boldsymbol{\pi_c} \neq \boldsymbol{p_c}$, expression (5) holds if p_w and p_ℓ are replaced with π_w and π_ℓ, provided the winner's true share of valid votes, $\pi_w/(\pi_w + \pi_\ell)$, is a bit greater than $(s_w + 0.5)/2$.[14] Otherwise—if the actual margin is not more than half the reported margin—the ASN is undefined: The audit may have a positive probability of continuing indefinitely, even if the reported outcome is correct. Of course, the auditors may elect to perform a full hand count at any time, and a full hand count might be prudent if the true margin is half or less than half of the reported margin, even if the outcome is correct.

7 Plurality contests with $C > 2$ candidates

Many contests have more than two candidates. Vote-for-one plurality contests in which one candidate receives a majority of the vote can be audited using the method described in the previous section by "pooling" votes cast for all the reported losers as if they were votes for a single hypothetical loser ℓ. Here, we present a more flexible and efficient approach that works for k-winner plurality contests with $k \geq 1$.

Consider each pair of an apparent winner and an apparent loser, (w, ℓ), $w \in \mathscr{W}$, $\ell \in \mathscr{L}$. The approach tests the $k(C-k)$ null hypotheses $\{\pi_w \leq \pi_\ell\}$, using the same sample but different test statistics $\{T_{w\ell}\}$ Each pairwise comparison relies only on valid votes cast for the candidates w and ℓ. The multipliers that update $T_{w\ell}$ after each ballot is drawn depend only on the reported votes for those two candidates.

In the most general case we consider, each ballot may have valid votes for zero or more reported winners, and for zero or more reported losers. Candidate w really beat candidate ℓ if and only if more than half the ballots that show a vote for either w or ℓ—but not both—show votes for candidate w. (Ballots that show votes for both w and ℓ, or for neither w nor ℓ, favor neither candidate.) Testing whether w really beat ℓ won therefore amounts to testing whether candidate ℓ got half or more of the votes on such ballots.

Let $s_{w\ell} \equiv s_w/(s_w + s_\ell)$ be the fraction of votes w was reported to have received among ballots reported to show a vote for w or ℓ or both. The value of $s_{w\ell}$ can be calculated from standard reported election results, whereas the fraction of ballots reporting votes for w but not ℓ cannot be. Hence, our test for the pair (w, ℓ) should not require knowing that fraction.

Suppose that w reportedly beat ℓ, so that $s_{w\ell} > 0.5$, and suppose that a fraction s of ballots reportedly showed votes for both w and ℓ. Then the fraction of ballots reported to have votes for w but not ℓ among those that show votes for w or ℓ but not both is

$$\frac{s_w - s}{s_w + s_\ell - 2s} > s_{w\ell}. \tag{8}$$

That is, the reported margin for w among ballots with votes for exactly one of w and ℓ is larger than the reported margin among ballots reported to show votes for one or

[13] This estimate is derived from expression (3:57) in Wald [2004], setting B and $L(\theta)$) to 0. (Expression (4.8) in Wald [1945] is equivalent.) The additional term $z_w/2$ allows for the fact that the final value of T ordinarily exceeds $1/\alpha$.

[14] The denominator of equation (5) vanishes if the winner's true share of valid votes equals $\ln(1-x)/(\ln(1-x) - \ln(1+x))$. This value exceeds $(s_w + 0.5)/2$ by less than 0.001 for $s_w \leq 0.642$.

Winner's	Quantiles						
True Share	25^{th}	50^{th}	75^{th}	90^{th}	99^{th}	Mean	ASN
70%	12	22	38	60	131	30	30
65%	23	38	66	108	236	53	53
60%	49	84	149	244	538	119	119
58%	77	131	231	381	840	184	185
55%	193	332	587	974	2,157	469	469
54%	301	518	916	1,520	3,366	730	731
53%	531	914	1,619	2,700	5,980	1,294	1,295
52%	1,188	2,051	3637	6,053	13,455	2,900	2,902
51%	4,725	8,157	14,486	24,149	53,640	11,556	11,562
50.5%	18,839	32,547	57,838	96,411	214,491	46,126	46,150

Table 1: Estimated means and percentiles of the number of ballots with valid votes to inspect for 10% risk limit using BRAVO, as a function of the winner's share of vote (estimated using 10^7 replications), as well as Wald's ASN.

both of w and ℓ. Hence, it is conservative to use $s_{w\ell}$ as the basis for the multiplier in the test. This leads to the complete BRAVO procedure:

1. Set $m = 0$ and set $T_{w\ell} = 1$ for all $w \in \mathscr{W}$ and $\ell \in \mathscr{L}$.

2. Draw a ballot uniformly at random with replacement from those cast in the contest and increment m.

3. If the ballot shows a valid vote for a reported winner w, then for each ℓ in $\mathscr{L}$ that did not receive a valid vote on that ballot multiply $T_{w\ell}$ by $s_{w\ell}/0.5$. Repeat for all such w.

4. If the ballot shows a valid vote for a reported loser ℓ, then for each w in $\mathscr{W}$ that did not receive a valid vote on that ballot multiply $T_{w\ell}$ by $(1 - s_{w\ell})/0.5$. Repeat for all such ℓ.

5. If any $T_{w\ell} \geq 1/\alpha$, reject the corresponding null hypothesis for each such $T_{w\ell}$. Once a null hypothesis is rejected, do not update its $T_{w\ell}$ after subsequent draws.

6. If all null hypotheses have been rejected, stop the audit: The reported results stand. Otherwise, if $m < M$, return to step 2.

7. Perform a full hand count; the results of the hand count replace the reported results.

This method limits the overall risk to α: Stopping short of a full hand count is an error only if at least one of the null hypotheses is in fact true. The audit stops only if all of the null hypotheses are rejected. Consider the set of null hypotheses that are true. The chance BRAVO erroneously rejects *all* of those and stops without a full hand count is at most the smallest chance of erroneously rejecting any of them 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.

For any given risk limit, the expected number of draws to confirm a correct outcome using BRAVO generally depends primarily upon the smallest margin of decision—the difference in vote shares between the winner with the smallest vote share and the loser with the largest vote share.[15] Call these candidates w^* and ℓ^*. In the vote-for-one case, if no other margin of decision is very close to the smallest one (between the reported winner and runner-up), then the expected number of ballots with valid votes to inspect is very close to the expression for the ASN above, setting $p_w = p_{w^*}$, $p_\ell = p_{\ell^*}$, $s_w = p_{w^*}/(p_{w^*} + p_{\ell^*})$, and $s_\ell = 1 - s_w$ —as if all ballots not cast for one of the two leading candidates were invalid.[16]

However, if one or more other margins of decision are close to the smallest one, then the expected number of ballots may be substantially larger, as it becomes harder to reject all the pairwise null hypotheses at once. For instance, in a three-candidate contest where the candidate vote shares are 40%, 30%, and 30%, the average sample size (determined by simulation with 10^6 trials) is approximately 433 ballots. This number is modest, but about 31% larger than the ASN formula indicates;

[15] This generalization holds (with the caveat described next) when discrepancies between reported and actual vote shares are small relative to the differences among candidates' actual vote shares; in contests that allow a voter to vote for more than one candidate, it also may depend on the fraction of ballots that show votes for both members of each (winner, loser) pair.

[16] For instance, consider a three-candidate contest where the vote shares are 49.5%, 40.5%, and 10%. The apparent winner's share of the top-two vote is $p_{w^*}/(p_{w^*} + p_{\ell^*}) = 0.495/(0.495 + 0.405) = 0.55$. The average number of ballots inspected, as determined by simulation with 10^6 trials, is about 521. This is essentially equal to the value of ASN (expression (5)) with $p_w = 0.495$, $p_\ell = 0.405$, $s_w = 0.55$, and $s_\ell = 0.45$, or 10/9 the value of ASN with $s_w = p_w = 0.55$ and $w_\ell = p_\ell = 0.45$.

however, simulating the workload is still quite tractable, even for rather small margins. In contests that allow each voter to select more than one candidate, the situation is even more complicated, and may depend on the number of ballots with valid votes for each (winner, loser) pair. Then, even simulating the workload becomes knotty, because it requires assumptions about those "overlaps," and there is little data to support the assumptions. However, these details affect only the workload, not the risk limit: BRAVO does not require any assumptions about the margins or the overlap to limit the risk to α rigorously.

8 Historical examples

To explore the potential applicability of BRAVO, we examined state-level[17] vote totals from the five U.S. presidential elections from 1992 through 2008. In practice, neither BRAVO nor any other risk-limiting audit method could have been applied in all states in all these elections: Some states have used (and some states now use) voting systems that do not provide a voter-verifiable audit trail. Nevertheless, it seems worthwhile to consider the empirical distribution of reported vote shares in all these states in recent elections.

We extend the analysis to 1992 because Ross Perot's candidacy that year offers the most interesting recent example of a viable third-party candidacy; as we expected, it had little impact on the results.[18] We use vote counts from Dave Leip's Atlas of U.S. Presidential Elections (`uselectionatlas.org`), including counts of invalid ballots where available.[19] We estimated the expected sample sizes needed to confirm the reported outcomes at a risk limit of 10% by simulation. The simulations assume that the reported vote shares are correct. Each simulation estimates the expected number of ballots required to confirm the plurality winner in one state in one election, running BRAVO 10^6 times for each such contest.

Of the 255 state presidential contests in this period, we set aside 23 that had margins smaller than 2 percent. For a 2-percent margin, the expected number of ballots to inspect is over 11,000 (see Table 1 for winner's share 51%). We think that many states would find it onerous to individually sample tens of thousands of ballots, especially with no clear expectation of when they can stop. Eight of these 23 contests had margins under 0.25 percent, below the threshold for an automatic recount in some states; hand counts (where feasible) arguably would be the most efficient and trustworthy means of confirming those outcomes. In the intermediate cases, alternatives to BRAVO, such as ballot-level comparison audits, might be preferable.

In the remaining 232 state-level presidential contests (91% of the contests under examination), the total expected number of ballots to inspect in order to confirm all outcomes was approximately 225,000, out of almost 512 million ballots cast in those contests. In 49 cases, the expected sample size was under 100 ballots; in 179 cases (70% of all contests under examination), the expected sample size was under 1,000 ballots. The median expected sample size (treating the 23 closest contests as if they required infinitely large samples) was about 307 ballots.

Broadly speaking, we think BRAVO at a 10% risk limit would not be onerous for most states in most elections. However, logistical challenges of statewide audits and alternative methods for the hardest cases require more attention than we offer here.

9 Logistics

9.1 Drawing a random sample

Drawing a simple random sample of ballots is not as straightforward as drawing marbles from an urn. Even if it were possible physically to mix the ballots, it is imprudent. A better approach starts by implicitly enumerating the ballots using a *ballot manifest*. A simple ballot manifest lists the physical containers of ballots in which the ballots are stored and how many ballots are in each container. The auditors can sequentially generate random numbers, convert each one to a ballot number (from 1 to the total number of ballots in the election), and convert each number to a particular ballot, for instance, "the 142nd ballot in box 41." If the contests under audit are on almost every ballot, we can simply treat ballots that do not contain the contest as having no valid vote, but if the contests under audit are on only a small fraction of the ballots, many draws will yield useless ballots. Retrieving those ballots can be an expensive waste of time.

If the contests under audit appear only on some of the ballots in the election, we can reduce the number of useless draws if we can construct a ballot manifest specific to those elections from the comprehensive manifest; see, e.g., Lindeman and Stark [2012].

[17] We treat the District of Columbia as a state. For purposes of this analysis, we do not consider the Maine and Nebraska electors chosen at the congressional district level.

[18] By way of example: In Maine, Perot narrowly edged out George H.W. Bush for second place: The vote shares were Bill Clinton 38.77%; Perot 30.44%; Bush 30.39%; other candidates 0.41%. The resulting expected sample size was about 610 ballots—more than the 470 ballots one would expect if only the Clinton-Perot comparison were considered, but not burdensome. In most states the impact was far smaller, essentially undetectable.

[19] Invalid ballot counts were not available for 1992 and 1996. Also, in 2008, Connecticut reported fewer ballots cast than presidential votes counted; we assumed that no invalid votes were cast. Imputing invalid vote counts would not materially affect the results.

Unless the ballot manifest uses specific ballot identifiers, the audit needs a trustworthy way to identify a particular physical ballot in the sample. Innocent unpredictable errors in identifying, say, the 142nd ballot in a box should not, in principle, bias the results. However, reasonable suspicion that systematic error exists or that the auditors can exercise discretion over which ballots are selected undermines the value of the audit. Therefore, care should be taken to specify and follow credible procedures.

If the ballot manifest has errors, the chance of drawing each ballot will not be equal. Some ballots might have no chance at all of being drawn, and the audit might attempt to draw ballots that do not in fact exist—or that exist but are not where the manifest claims they are. As a result, the sampling distribution of the test statistics $\{T_{w\ell}\}$ are not what they were assumed to be, which could make the actual risk limit larger than claimed. Fortunately, there is an easy remedy.

Bañuelos and Stark [2012] show that if there is an upper bound on the number of ballots, a simple modification to the procedure makes the audit *more* conservative if the manifest has errors. That is, the actual risk limit will be even smaller than the claimed risk limit. The modification is simple: If the upper bound on the number of ballots is larger than the total listed in the manifest, create a fictitious group of ballots that contains the extras, and append it to the manifest. In each draw in the audit, sample uniformly from 1 to the upper bound on the number of ballots. If the ballot drawn cannot be found—either because it is in the fictitious group or because the manifest lists more ballots in a batch than are actually there—pretend that a ballot was found, and that the ballot showed a valid vote for every loser ℓ.[20] Perhaps surprisingly, it is not necessary to adjust the margin to account for missing ballots or ballots not listed in the manifest, only to treat the ballots actually selected in the course of the audit in the most pessimistic way.

9.2 Group sequential sampling versus item-by-item audits

For practical reasons, a jurisdiction may prefer to retrieve some number of ballots and inspect them together, rather than retrieving and inspecting one ballot at a time. For instance, it may be convenient to select, say, 100 ballots, sort the selections based on the containers in which the ballots are stored, and retrieve those ballots. If any such sorting occurs—if ballots are audited in a different order than the order in which they were randomly selected—then all the ballots in a selected group *must* be audited. Using groups in this manner increases the expected work, because the audit cannot end in the middle of inspecting a group.[21] However, as long as sorted groups are audited completely, using group sampling does not compromise the risk limit α.

Moreover, the group size can be varied at will throughout the audit. For instance, if a jurisdiction can audit additional groups rather easily and would like to limit the amount of ballots it has to inspect, it might calculate the ASN and begin the audit by auditing half that many ballots, which provides a non-trivial (on the order of 25%) chance of completing the audit in the first group. Conversely, if limiting the number of stages is more important than limiting the number of ballots counted, the jurisdiction might use the ASN, or even some multiple of the ASN, as the sample size in the first group. The jurisdiction can then adjust subsequent group sizes based on how far the test statistics are from $1/\alpha$ and how highly it values limiting the number of ballots inspected versus limiting the number of groups. For instance, in the two-candidate case, the conditional ASN to confirm the reported outcome given that the results so far yield a test statistic $T = T^*$, assuming that the reported vote shares are correct, is derived from expression (5) by replacing $\ln(1/\alpha)$ with $\ln((1/\alpha)/T^*)$ in the numerator. The jurisdiction can use this conditional ASN to help it decide how many ballots to sample in the next group.[22]

9.3 Coordinating audits across multiple jurisdictions

Ballot-polling audits show particular promise in large, multi-jurisdictional elections because the expected number of ballots to inspect often will be a tiny fraction of the ballots in the election, and because the work can be divided among many election officials. However, they also pose distinctive logistical challenges. It is one matter for a particular jurisdiction to conduct a sequential audit ballot-by-ballot, or to choose a convenient alternative. It is another matter for dozens or hundreds of local election officials to participate in a ballot-by-ballot sequential audit. Given modern communications, a coordinated ballot-by-ballot audit could be feasible: At each step, a ballot would be randomly selected from all the ballots in the contest, regardless of jurisdiction, and the appropriate officials would locate and inspect that ballot.

[20] The rules of the election might allow a ballot to have valid votes for fewer candidates than there are losers ℓ; nonetheless, for this method to result in a conservative procedure, the auditors should pretend that the ballot gives a valid vote to every loser.

[21] The test statistic may exceed $1/\alpha$ in the course of a group but drop below $1/\alpha$ by the end of the group. In that case, one or more additional groups must be counted before the audit ends.

[22] For finer control, it is possible to estimate quantiles of the expected number of ballots to sample: The jurisdiction can select a sample size that is expected to provide, say, a 90% chance of completing the audit if the reported vote shares are correct. We do not explore these details here.

Even if this approach is feasible, it is likely to become tedious for all but the smallest audits. Therefore, multi-jurisdictional audits are likely to be conducted by group sampling rather than item-by-item.

If each jurisdiction can provide its ballot manifests to a central election authority in a machine-readable common format, the central authority can produce a contest-wide master manifest, and then use this manifest to conduct the sampling. However, if it is not immediately practical for all jurisdictions to produce ballot manifests in a common format, it suffices for each jurisdiction to provide the number of ballots enumerated in the manifest. Then, if desired, the central authority can conduct the sample, in each case telling the appropriate jurisdiction to inspect its *x*th ballot. A more complex alternative, which allows jurisdictions to draw their own samples, is a two-stage sample: The central authority draws a sample to determine how many ballots each jurisdiction should inspect (which will, in general, vary across jurisdictions). Each county then separately samples and inspects as many ballots as required. Regardless of the method used, the jurisdictions report their results to the central authority, which combines them to determine whether the audit can stop. Each of these sampling methods can be repeated as often as necessary.

Auditors in multi-jurisdictional audits are likely to place a higher priority on limiting the number of groups to be audited than limiting the number of ballots to inspect. This is true because of the challenges of coordinating an unpredictable number of groups across multiple jurisdictions, and also because widely distributing the auditing work reduces the efficiency bind: Many hands make work light. For instance, if one jurisdiction conducts an audit with an ASN of 500 ballots, it might prefer to audit 500 (or fewer) ballots in the initial group, rather than to audit 1000 ballots or more simply to increase its chance of finishing in one group. In contrast, if the State of California conducts a statewide audit with an ASN of 500 ballots, it might prefer an initial sample of 1000 ballots or more, knowing that the work will be distributed across 58 counties. Again, the number of ballots to be audited in each group can be informed by the audit results for previous groups, to reduce the chance of needing to audit additional groups, or to reduce the expected number of ballots to audit.

9.4 Pairwise comparisons versus "pooling"

Imagine auditing a vote-for-one plurality contest. If the reported winner apparently received more than half the votes, there are at least two approaches we could take to auditing: (i) "pool" some or all the losing candidates and test the null hypothesis that each "pooled" group received more votes than the reported winner, and (ii) test the $C-1$ hypotheses that each of the $C-1$ apparent losers got more votes than the reported winner. We show in this section that the latter approach is more efficient; the proof also applies to the general case BRAVO considers: plurality contests with k winners in which each voter may select zero or more candidates.

Suppose that the reported results are correct. We will show that, for any (winner, loser) pair (w,ℓ) and any number n of draws, $\Pr\{T_{w\ell} \geq 1/\alpha\}$ is larger than $\Pr\{T_{wl} \geq 1/\alpha\}$, where l is any composite "pooled" candidate consisting of ℓ and any subset of the other losers. Since we stop the audit only when all the test statistics are greater than $1/\alpha$, it follows that the expected sample size is smaller if candidates are not pooled.

Recall that s_w is w's share of the valid ballots and s_ℓ is the loser's share. Let s_m be the combined share of any other losers with which we are considering pooling ℓ. To pool ℓ with m, we require $s_w > s_\ell + s_m$; otherwise, w was not reported to get more votes than the pooled loser. Consider the jth draw. Let I_{wj} denote the event that the ballot drawn shows a vote for w, $I_{\ell j}$ the event that the ballot shows a vote for ℓ, and I_{mj} the event that the ballot shows a vote for some other loser we are grouping with. Then $I_{wj} + I_{\ell j} + I_{mj} \leq 1$, $\Pr(I_{wj} = 1) = s_w$, and so on.

We focus on a single draw (since the draws are independent and identically distributed) and condition on the event that the ballot drawn shows a valid vote; otherwise, it cannot help, whether we pool or not. The event $T \geq 1/\alpha$ is the same as the event $\ln T \geq \ln 1/\alpha$. The random variable $\ln T$ is the sum of independent, identically distributed increments—a random walk with drift. Without pooling, the increment to $\ln T$ from the jth draw is

$$Z_{j\ell} = I_w \ln 2 \frac{s_w}{s_w + s_\ell} + I_\ell \ln 2 \frac{s_\ell}{s_w + s_\ell}. \tag{9}$$

With grouping, the increment is

$$Z_{jl} = I_w \ln 2 \frac{s_w}{s_w + s_\ell + s_m} + (I_\ell + I_m) \ln 2 \frac{s_m + s_\ell}{s_w + s_\ell + s_m}. \tag{10}$$

A result in Wald [1945] implies that the expected number of observations necessary to reach a decision is $1/\alpha$ divided by the expected increment to the test statistic: The larger the expected increment, the smaller the expected number of draws. Hence, it suffices to show that the expected value of $D \equiv Z_{j\ell} - Z_{jl}$, the difference in the expected increments without and with grouping, is nonnegative. Observe:

$$\begin{aligned}
D &\equiv s_w \ln \frac{s_w + s_\ell + s_m}{s_w + s_\ell} \\
&\quad + s_\ell \ln \frac{s_\ell (s_w + s_\ell + s_m)}{(s_w + s_\ell)(s_m + s_\ell)} \\
&\quad - s_m \ln 2 \frac{s_m + s_\ell}{s_w + s_\ell + s_m} \\
&= s_w \ln \frac{s_w + s_\ell + s_m}{s_w + s_\ell} \\
&\quad + \ln \frac{s_\ell^{s_\ell} (s_w + s_\ell + s_m)^{s_\ell + s_m}}{2^{s_m} (s_w + s_\ell)^{s_\ell} (s_m + s_\ell)^{s_m + s_\ell}}.
\end{aligned}$$

Since $\frac{s_\ell}{s_w + s_\ell} < \frac{1}{2}$,

$$\begin{aligned}
D &\geq s_w \ln \frac{s_w + s_\ell + s_m}{s_w + s_\ell} \\
&\quad + \ln \left[\frac{1}{2^{s_m + s_\ell}} \frac{(s_w + s_\ell + s_m)^{s_\ell + s_m}}{(s_m + s_\ell)^{s_m + s_\ell}} \right] \\
&= (s_w + s_\ell + s_m) \ln [s_w + s_\ell + s_m] \\
&\quad - s_w \ln [s_w + s_\ell] \\
&\quad - (s_\ell + s_m) \ln [2 (s_m + s_\ell)]
\end{aligned}$$

Finally since $s_w + s_\ell + s_m > s_w + s_\ell$, and $s_w > s_\ell + s_m$,

$$\begin{aligned}
D &\geq (s_w + s_\ell + s_m) \ln [s_w + s_\ell + s_m] \\
&\quad - s_w \ln [s_w + s_\ell + s_m] \\
&\quad - (s_\ell + s_m) \ln [s_w + s_\ell + s_m] \\
&= 0. \square
\end{aligned}$$

10 Discussion

BRAVO provides a way to perform a risk-limiting audit of majority, super-majority, and plurality contests, including contests with more than one winner and contests in which voters may cast votes for more than one candidate. BRAVO places minimal demands on the voting system: It requires the reported contest results, an audit trail, and a ballot manifest that explains how the ballots are stored, so that ballots can be selected at random. In contrast, comparison audits require detailed reports from the voting system that are not produced in machine-readable form by current vote tabulation systems.

The number of ballots that must be audited using BRAVO when the reported results are correct can be quite small, and that workload is distributed over all the jurisdictions involved in the contest. That makes BRAVO immediately practical for jurisdictions that have an audit trail, for contests with margins down to a few percent. In particular, the median expected sample size for states' presidential contests from 1992 through 2008 is about 307 ballots.

BRAVO does not check the accuracy of the voting system, only the correctness of the electoral outcome. The voting system could get the right outcome through fortuitous cancellation of errors, which a comparison risk-limiting audit might detect. The workload for BRAVO becomes prohibitive when the margin is small; in such cases, ballot-level or batch-level comparison audits might be preferable, despite their higher set-up costs.

11 Acknowledgments

We are grateful to Jennie Bretschneider, Joseph Lorenzo Hall, Ronald L. Rivest, and Pamela Smith for helpful comments.

References

Bañuelos, J. and Stark, P. (2012). Limiting risk by turning manifest phantoms into evil zombies. Technical report, arXiv.org. `http://arxiv.org/abs/1207.3413`. Retrieved 17 July 2012.

Benaloh, J., Jones, D., Lazarus, E., Lindeman, M., and Stark, P. (2011). SOBA: Secrecy-preserving observable ballot-level audits. In *Proceedings of the 2011 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '11)*. USENIX.

Calandrino, J., Halderman, J., and Felten, E. (2007). Machine-assisted election auditing. In *Proceedings of the 2007 USENIX/ACCURATE Electronic Voting Technology Workshop (EVT 07)*. USENIX.

Checkoway, S., Sarwate, A., and Shacham, H. (2010). Single-ballot risk-limiting audits using convex optimization. In *Proceedings of the 2010 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '10)*. USENIX. `http://www.usenix.org/events/evtwote10/tech/full_papers/Checkoway.pdf`. Retrieved April 20, 2011.

ElectionAudits.org (2008). Principles and best practices for post-election audits. `http://www.electionaudits.org/files/best%20practices%20final_0.pdf`. Retrieved May 10, 2012.

Gibbons, J. D., Olkin, I., and Sobel, M. (1977). Selecting and Ordering Populations : A New Statistical Methodology. In *Selecting and Ordering Populations: A New Statistical Methodology*, chapter 6, pages 158–186. SIAM.

Hall, J., Miratrix, L., Stark, P., Briones, M., Ginnold, E., Oakley, F., Peaden, M., Pellerin, G., Stanionis, T., and Webber, T. (2009). Implementing risk-limiting post-election audits in California. In *Proc. 2009 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE '09)*, Montreal, Canada. USENIX.

Johnson, K. (2004). Election certification by statistical audit of voter-verified paper ballots. `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=640943`. Retrieved 6 March 2011.

Lindeman, M. and Stark, P. B. (2012). A gentle introduction to risk-limiting audits. *IEEE Security and Privacy*, page to appear.

McDaniel, P., Blaze, M., Vigna, G., and et al. (2007). EVEREST: Evaluation and Validation of Election-Related Equipment, Standards and Testing. Retrieved 17 March 2012.

Ng, H. K. T. and Panchapakesan, S. (2007). Is the Selected Multinomial Cell the Best? *Sequential Analysis*, 26(4):415–423.

Ramey, J. and Alam, K. (1979). A sequential procedure for selecting the most probable multinomial event. *Biometrika*, pages 171–173.

Rivest, R. and Shen, E. (2012). A Bayesian method for auditing elections. In *Proceedings of the 2012 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '12)*. USENIX.

Secretary of State, C. (2007). Top-to-bottom review. Retrieved 9 July 2012.

Simon, J. D. and O'Dell, B. (2006). An end to 'faith-based' voting: universal precinct-based handcount sampling to check computerized vote counts in federal and statewide elections. `http://electiondefensealliance.org/files/UPSEndFaithBasedVoting.pdf`. Retrieved July 6, 2012.

Sorentrue, J., Kam, D., and Bennett, G. (2012). Recount shows wrong winners declared in two Wellington election races. Retrieved 3 May 2012.

Stark, P. (2008a). Conservative statistical post-election audits. *Ann. Appl. Stat.*, 2:550–581.

Stark, P. (2008b). A sharper discrepancy measure for post-election audits. *Ann. Appl. Stat.*, 2:982–985.

Stark, P. (2009a). CAST: Canvass audits by sampling and testing. *IEEE Transactions on Information Forensics and Security, Special Issue on Electronic Voting*, 4:708–717.

Stark, P. (2009b). Efficient post-election audits of multiple contests: 2009 California tests. `http://ssrn.com/abstract=1443314`. 2009 Conference on Empirical Legal Studies.

Stark, P. (2009c). Risk-limiting post-election audits: *P*-values from common probability inequalities. *IEEE Transactions on Information Forensics and Security*, 4:1005–1014.

Stark, P. (2010). Super-simple simultaneous single-ballot risk-limiting audits. In *Proceedings of the 2010 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '10)*. USENIX.

Stark, P. B. and Wagner, D. A. (2012). Evidence-based elections. *IEEE Security and Privacy*, page to appear.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Stat.*, 16:117–186.

Wald, A. (2004). *Sequential Analysis*. Dover Publications, Mineola, New York.