Workload Estimates for Risk-Limiting Audits of Large Contests

Katherine McLaughlin and Philip B. Stark
Department of Statistics
University of California, Berkeley

5 June 2011

Abstract

We compare the expected number of ballots that must be counted by hand for two risk-limiting auditing methods, Canvass Audits by Sampling and Testing (CAST) and Kaplan-Markov (KM). The methods use different sampling designs to select batches of ballots to count by hand and different test statistics to decide when the audit can stop. The comparisons are based on the 2008 U.S. House of Representatives contests in California. The comparisons include hypothetical errors in the precinct vote totals, but errors are assumed to be small enough that the electoral outcomes are still correct. KM requires auditing fewer ballots than CAST. The workload for CAST can be reduced modestly by optimizing the number of precincts drawn from each county. Stratification by county is necessary for the practical implementation of risk-limiting audit methods in cross-jurisdictional contests. Workload can be reduced substantially, for both KM and CAST, by tallying ballots in batches smaller than precincts: Workload is roughly proportional to the average size of the batches. We discuss several methods to reduce batch sizes using current vote tabulation systems.

1 Introduction to Risk-Limiting Election Auditing

No method of tallying votes is invulnerable to error, including the electronic vote tabulation systems used in most U.S. elections. The exact vote totals are less important than the *electoral outcome*, the set of winners. For election integrity, we would like to be able to assess whether errors in the vote totals changed the electoral outcome. If there is a complete, accurate, durable audit trail, a full hand count of that trail would give the "true" electoral outcome (generally as a matter of law). In principle, we could check whether the *apparent outcome*—the electoral outcome that will become final unless something intervenes—agrees with the true outcome, by counting the entire audit trail by hand. This is clearly expensive and inefficient. The goal of statistical post-election vote tabulation audits is to ensure that, if the apparent outcome does not agree with the correct outcome, it is very likely to be corrected by the audit—and to provide that assurance with as little hand counting as possible.

To conduct an audit, hand tallies of the votes in *batches* of ballots are compared to the original machine tallies of the votes in those batches.[1] Audits can only check subtotals that the vote tabulation system reports. Current commercial vote tabulation systems only report subtotals for fairly large batches—precincts or precincts subdivided into ballots cast in person or by mail are typical. If vote tabulation systems reported subtotals for much smaller batches, that would greatly reduce the amount of hand counting required to audit to limit risk to a given level, as our simulations in Section 5 confirm.

Vote tabulation audits serve many purposes; here we are concerned with audits that provide statistical assurance that the outcome is correct. An audit is *risk-limiting* if and only if it has a known minimum probability of requiring a full manual count whenever the apparent outcome is wrong, thereby correcting the

[1] There are ways of performing election audits that do not require comparing original machine tallies to hand counts, for instance *ballot polling*, which uses a random sample of ballots to test the hypothesis that the apparent winner is not the true winner. See Stark [2011b].

apparent outcome. A risk-limiting audit ends either (1) with a full hand count, which replaces the apparent outcome if the outcome according to the hand count differs from the apparent outcome; or (2) without a full hand count, leaving the apparent outcome as the official outcome. In the last few years, risk-limiting audits have been widely acknowledged to be best practice [Lindeman et al., 2008]. They have been endorsed by the American Statistical Association, Common Cause, the League of Women Voters, Verified Voting, and Citizens for Election Integrity Minnesota, among others. A formal pilot of risk-limiting audits is required by California Assembly Bill 2023 [Saldaña, 2010], which became law in July, 2010. Colorado Revised Statutes §1-7-515 calls for risk-limiting audits by 2014. In May 2011, California and Colorado were awarded grants from the U.S. Election Assistance Commission to develop and implement practical procedures for risk-limiting audits.

Any method for risk-limiting audits has a large chance of leading to a full count when the apparent outcome is wrong. But different methods have different counting burdens when the apparent outcome is correct: An efficient method keeps the hand counting to a minimum when the apparent outcome is right. The workload depends on many things, including the margin, the sampling scheme, the sizes of auditable batches, the test statistic, and the number and nature of errors that the audit finds. This paper compares the workload required by two generic approaches to risk-limiting audits in hypothetical scenarios based on the 2008 California House of Representatives contests. Workload is defined to be the percentage of ballots cast that must be counted by hand before the audit stops.[2] In the hypothetical scenarios, the vote totals have only small errors, which we expect is typical when voting systems work properly and the reported outcome is correct. We examine how details of the sampling strategies (for instance, sampling fractions across strata) affect workload. We also study the savings that would be possible if vote-tabulation systems could report subtotals for smaller batches.

2 The Data

To date, there have been nine risk-limiting audits [Hall et al., 2009; Stark, 2009b; McBurnett, 2010], eight in California[3] and and one in Boulder County, Colorado; none has involved a multi-jurisdictional (multi-county) contest. A risk-limiting audit of a multi-jurisdictional contest requires cooperation and coordination among election officials in neighboring counties. Here, we estimate the workload for auditing multi-county contests using data from the 4 November 2008 general election in California.

That election included the state's 53 seats in the U.S. House of Representatives, making it ideal for studying large cross-jurisdictional contests. The districts are large, with over 130,000 ballots cast in each. Some congressional districts intersect as many as ten counties, while others are contained within a single county that also includes other congressional districts.

The data were obtained from the Statewide Database (SWDB) [Institute of Governmental Studies, University of California]. For the comparisons, we used data from 44 of the 53 congressional districts. We omitted seven districts because a candidate was running unopposed and two because the SWDB coding did not separate votes received by independent candidates from *undervotes* (ballots with no candidate selected in the contest) and *overvotes* (ballots with more than one candidate selected in the contest). Roughly half (23 of 44) of the districts cross county lines. There were between two and five candidates on the ballot in the 44 contests, from six different political parties; the number of ballots cast ranged from 130,337 to 386,707; and the margin (as a percentage of votes cast) ranged from 0.465% to 70.90%. The contests vary widely in size and in margin, as shown in Figure 1. Congressional District 4, represented by the point in the lower right corner, had a very small margin (0.465%, or 1800 ballots, which was fewer votes than cast in one precinct in the district), which makes it anomalous in many of the comparisons below.

[2] Other factors also affect cost, for instance, the number of batches that must be retrieved from storage. However, we expect the primary driver of audit cost to be the number of ballots that must be handled and tallied.

[3] One audit each in Monterey, Orange, and Santa Cruz Counties, two audits in Marin County, and three audits in Yolo County.

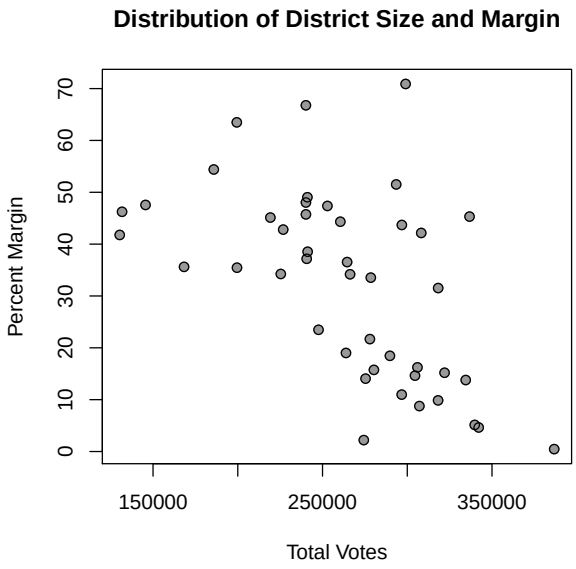

Figure 1: Votes cast and margins for 44 of the 2008 California U.S. House of Representatives contests.

3 Methods for Risk-Limiting Audits

Risk-limiting audits test the null hypothesis that the apparent outcome is incorrect by tallying the votes in batches of ballots selected at random. If the hand count gives strong evidence that the apparent outcome is not incorrect, the audit stops. Evidence is measured using P-values: The P-value is the largest chance that the audit would find the results it did find, if the apparent outcome were incorrect. Small P-values are strong evidence that the apparent outcome is correct. If we stop auditing only when the P-value is less than α, we guarantee that if the apparent outcome is wrong, the chance of a full hand count is at least $1-\alpha$. The risk limit is α.

We studied two methods for risk-limiting audits: Canvass Audits by Sampling and Testing (CAST) [Stark, 2009a] and Kaplan-Markov (KM) [Stark, 2009c,b]. Both methods have been used to audit contests entirely contained in a single jurisdiction [Stark, 2009b; Hall et al., 2009; Stark, 2009a]. These methods audit in stages. At stage s of the audit, the P-value P_s is calculated (conditional on the results of previous stages, for CAST). If P_s is less than a pre-specified threshold α_s, the audit stops and the apparent outcome is certified. Otherwise, the audit continues to the next stage. Eventually, either the audit finds strong evidence that the apparent outcome is right, or there has been a full hand count, which reveals the correct outcome. See Stark [2009c] for methods for computing P-values for election audits.

CAST can be used with any random sampling design. Here, we compare three: simple random sampling (CAST-SRS), stratified sampling using proportional allocation (CAST-PROP), and stratified sampling using optimal allocation (CAST-OPT). SRS is simplest in theory but in practice requires the most cooperation across counties. For both stratified methods, the strata are counties. Stratifying by county reduces the need for coordination across jurisdictions.

Errors in the machine counts can result in overstating one or more margins, understating one or more margins, or they can be neutral. In the version of CAST we test, the decision to stop the audit is based on the observed overstatements of the margins between the apparent winner and every apparent loser. If the sample sizes at each stage are set appropriately, the stopping rule $P_s < \alpha_s$ for CAST amounts to comparing the maximum (across audited batches) of a measure of overstatement error to a threshold t. Here, we use *taint* as the measure of overstatement error. The taint T_j of batch j is the largest relative overstatement

of any margin in that batch divided by the maximum possible relative overstatement of any margin in that batch: It is the actual error expressed as a fraction of the maximum possible error. See Stark [2009c].

In CAST, the constants $\{\alpha_s\}$ and t affect the expected workload. The sampling design matters as well, as we shall see. Suppose there are N batches of ballots in the contest, of which N_c are in county (stratum) c, for $c \in \{1, \ldots, C\}$. Let b_i be the number of ballots in batch i, for $i \in \{1, \ldots, N\}$. A SRS of n batches is equally likely to select every subset of n of the N batches; each of the N precincts has chance n/N of being selected. Thus, the expected number of ballots to be audited is $\frac{n}{N} \sum_{i=1}^{N} b_i$.

CAST-PROP and CAST-OPT use independent simple random samples from each of the C strata, but the sample size n_c for stratum c is chosen differently for the two methods. In CAST-PROP, n_c is proportional to the number of batches in stratum c. For example, if 10 precincts will be selected from a district with strata of sizes 20, 40, and 40 batches respectively, CAST-PROP will select two from the first stratum and four from each of the other two. CAST-OPT finds $\{n_c\}$ to minimize the P-value if the observed maximum taint were t, subject to the constraint that the total sample size $\sum_{c=1}^{C} n_c = n$. For both CAST-PROP and CAST-OPT, the expected number of ballots in the sample from stratum c is $\frac{n_c}{N_c} \sum_{i=1}^{N_c} b_{ic}$, where b_{ic} is the number of ballots cast in batch i of stratum c. The total expected number of ballots sampled across all strata is thus $\sum_{c=1}^{C} \frac{n_c}{N_c} \sum_{i=1}^{N_c} b_{ic}$. When the distribution of error bounds differs across strata, CAST-PROP and CAST-OPT can require smaller samples than CAST-SRS. CAST-OPT never requires a larger sample than CAST-PROP.

KM relies on PPEB (probability proportional to error bound) sampling [Aslam et al., 2007; Stark, 2009c]. PPEB sampling draws n times independently from the N batches. In each draw, the probability of selecting batch p is proportional to a bound u_p on the error that batch p can hold: Batches that can hold more error are more likely to be selected than batches that can hold less error. PPEB can be used with CAST, but the resulting method is less efficient than KM. The advantage of CAST is that it permits simpler, more transparent sampling schemes to be used. Because PPEB draws are independent, the same batch can be selected repeatedly. When this happens, the batch does not need to be tallied by hand again, but the taint of the batch enters the calculation as many times as the batch is selected. The KM P-value is

$$P_s = \prod_{j=1}^{s} \frac{1 - \frac{1}{U}}{1 - T_j},$$

where T_j is the taint of the batch selected in the jth draw and $U = \sum_{p=1}^{N} u_p$. In KM, $\alpha_s = \alpha$. If $P_s < \alpha$, the audit stops. Otherwise, s is incremented, another batch is selected by PPEB and counted by hand (if it has not been counted already), and the P-value is calculated again. This process continues until either $P_s < \alpha$ or a full hand count has been conducted. KM relies on PPEB sampling across all jurisdictions, although we are developing methods for auditing using stratified PPEB sampling, which would reduce the need for coordination among counties.

We use taint to measure miscount in both CAST and KM to compare their workloads in identical hypothetical scenarios. By controlling the taint we can compare how the methods perform with different amounts of observed vote-tabulation error. The `elec` package [Miratrix, 2009] gives `R` implementations of CAST-SRS and KM. Theory and algorithms for CAST-OPT are given in Higgins et al. [2011].

4 Workloads for CAST and KM

We compare CAST and KM in hypothetical scenarios based on the 44 U.S. House of Representatives contests in California in 2008 described in Section 2. All the calculations described below use risk limit $\alpha = 10\%$. CAST tests used single-stage designs: If too much error is observed in the first sample, the audit immediately requires a full hand count, rather than escalating more gradually. Single-stage designs minimize workload when there are only small errors, which we believe is common unless something has gone seriously wrong—which would justify a full hand count.

We performed two sets of comparisons for each method. In the first set, no batch in the sample has error; that is, the taint of every batch is 0. In the second set, every batch in the sample has taint 0.01. It is

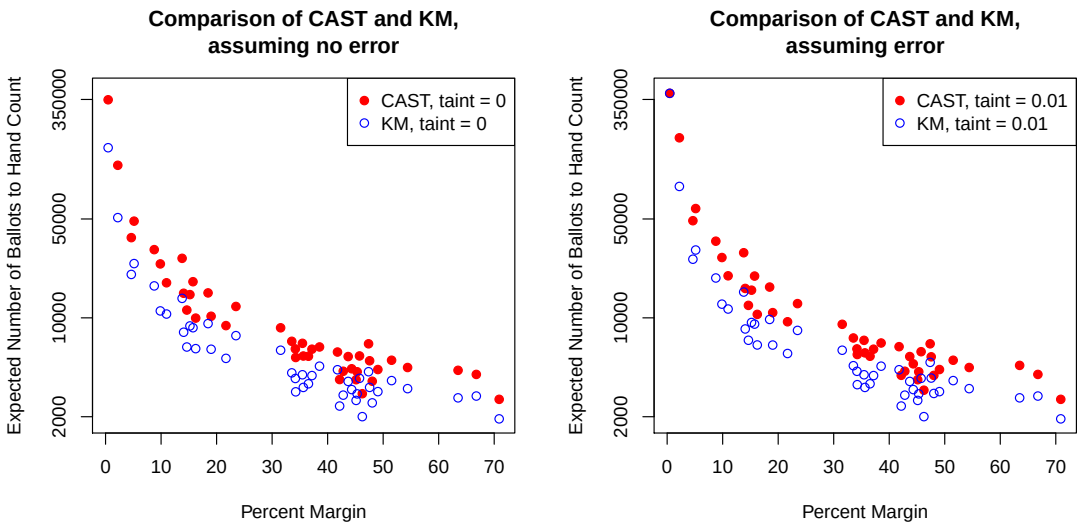

Figure 2: Expected number of ballots that must be audited for CAST-SRS and KM to attain a risk limit of 10% in 44 California contests, in two hypothetical error scenarios. CAST is tuned to require additional auditing only if the observed maximum taint exceeds zero (in the first error scenario) or exceeds 0.01 (in the second scenario); KM adapts automatically to the observed errors. Left: Expected ballots if the audit finds no error. Right: Expected ballots if the audit finds taint of 0.01 in every batch.

unrealistic (and generally pessimistic) to assume that every batch has taint 0.01. In a contest for which the apparent outcome is correct, we would expect some taints to be positive and some negative, but most to be zero. The expected numbers of ballots to hand count are nearly identical in the two error scenarios, so this deliberate pessimism does not have much effect on the workload estimates.

In both error scenarios, CAST is calibrated as if we knew ahead of time what the observed maximum taint would be. That is, in comparisons with taint 0, the CAST parameters are set to require a full hand count if any observed taint is greater than 0, and in comparisons with taint 0.01, the CAST parameters are set to require a full hand count if any observed taint is greater than 0.01. In contrast, KM is not "tuned" to any assumed level of error—it adapts to the observed error. This biases the comparison in favor of CAST. Moreover, CAST pays attention only to the largest error, while all errors matter in KM, so assuming that all the observed errors are equal also biases the comparison in favor of CAST. Despite these biases, KM does better.

Figure 2 shows the expected number of ballots to be hand counted for CAST-SRS and KM in both error scenarios, for each of the 44 contests, as a function of the margin in those contests (with margin measured as a percentage of votes cast). Despite the biases in favor of CAST, the workload for KM is lower than that for CAST except for the District 4 comparison with taint 0.01. In that contest, the margin (0.465%) is so small compared to the taint that both methods require essentially a full hand count.

To assess how workload depends systematically on features of the contests, we plot the ratio of the expected number of ballots to audit for each district:

$$\frac{\mathbb{E}\ (\text{ballots to audit for CAST})}{\mathbb{E}\ (\text{ballots to audit for KM})}.$$

For each district, we find the ratio in hypothetical scenarios in which the audit finds no error and in which the audit finds taint 0.01 in every batch. The results, plotted in Figure 3, demonstrate several things. First consider the plot with percent margin as the independent variable. Ratios range from approximately 1 to about 2.4: there were contests for which CAST required auditing more than twice as many ballots as

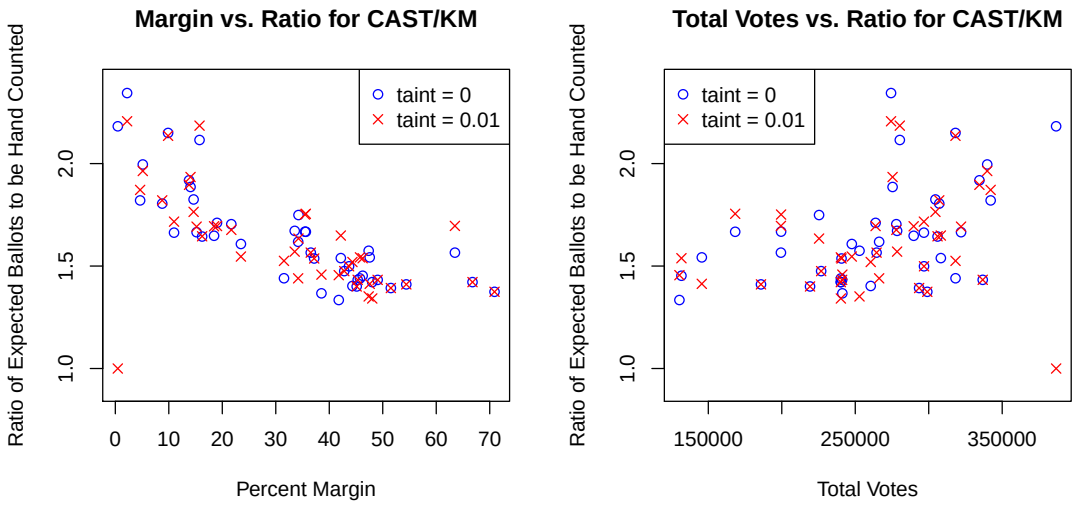

Figure 3: Ratio of expected number of ballots to be audited for CAST-SRS to expected number of ballots to be audited for KM, as a function of percent margin (left panel) and votes cast (right panel), for a risk limit of 10%, for 44 U.S. House of Representatives contests in California. Each plot shows two error scenarios: the audit finds no error, and the audit finds taint of 0.01 in every batch. CAST-SRS is tuned to require additional counting only if the observed error exceeds the error in the hypothetical scenarios, biasing the comparison in favor of CAST. District 4 is anomalous (the points on the far left of the left plot and on the far right of the right plot.

KM to certify the apparent outcome. KM seems to do particularly well when the margin is smaller. The level of error we are considering does not affect the expected workload much: The two curves overlap.

The second plot shows that KM generally required less auditing than CAST-SRS in larger contests. This may not generalize to other contests because because there was a negative association between margin and contest size for the 44 contests considered here (Figure 1)—larger contests tended to have smaller margins. This trend could result from confounding. Overall, these comparisons corroborate findings [Stark, 2009d,b] for smaller races and in theory: KM requires auditing fewer ballots to certify the apparent outcome than CAST-SRS does when the apparent outcome is correct.

Under California Elections Code §15360, counties draw their audit samples independently, which yields a stratified simple random sample of precincts. Of the 44 districts considered, 23 intersect at least two counties. Both KM and CAST-SRS require sampling from contests as a whole, coordinated across jurisdictions. Risk-limiting audits that use stratified sampling can reduce the need for coordination among jurisdictions. We consider CAST applied to stratified simple random samples using two rules for determining how many precincts to draw from each jurisdiction, CAST-PROP and CAST-OPT, described in Section 3.[4] We compare the expected workloads of these four methods in 20 of the 23 districts.[5]

For each method, the expected number of ballots to hand count was calculated assuming all audited batches were error-free and assuming that all audited batches had a taint of 0.01. As before, the parameters in the CAST methods were set as if we knew ahead of time what the observed errors would be: In the first scenario, CAST is calibrated to require a full hand count if any batch has taint greater than 0, and in the second scenario, CAST is set to require a full hand count if any batch has taint greater than 0.01. The

[4] Adapting KM to use stratified samples is a topic of current research. When a district only contains one stratum, CAST-SRS, CAST-PROP, and CAST-OPT are identical.

[5] We exclude three districts because a combination of a small margin and many strata (≥ 5) made it computationally infeasible to find the optimal stratified sample. We are developing faster algorithms.

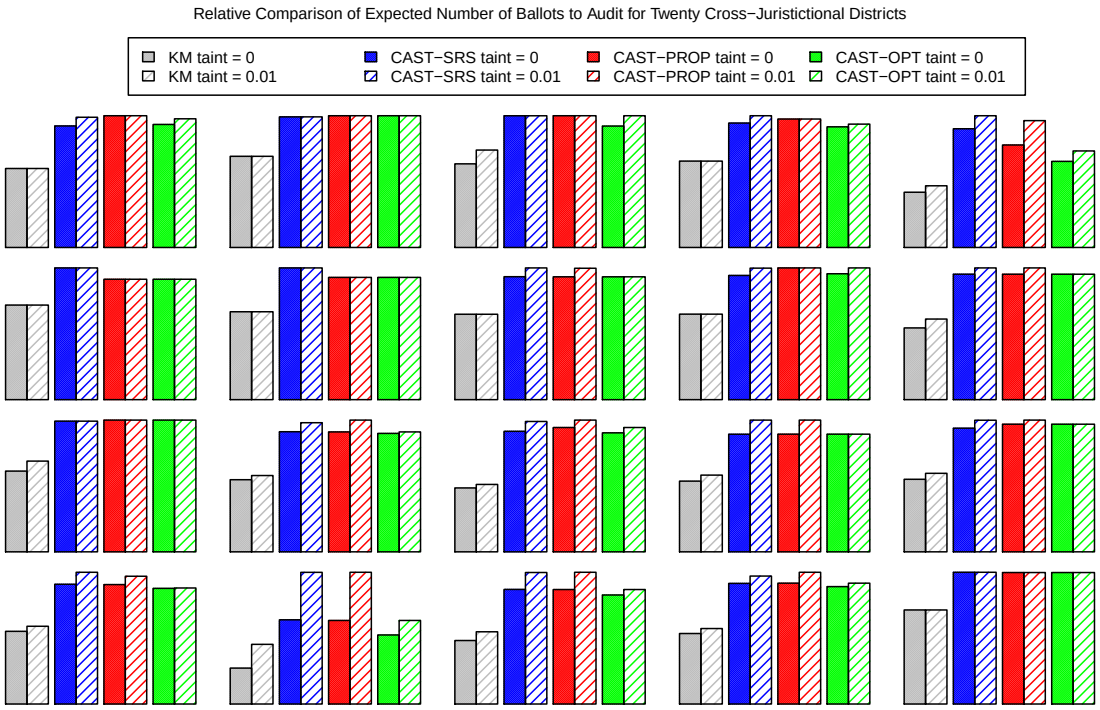

Figure 4: Estimated expected number of ballots to hand count for risk limiting audits using KM (gray), CAST-SRS (blue), CAST-PROP (red), and CAST-OPT (green) at risk limit 10% for the 20 contests that cross jurisdictions. Expected numbers if all observed taints are zero are shown as solid bars. Expected numbers if all observed taints are 0.01 are shown as striped bars. The parameters of the CAST methods are set as if we knew ahead of time what the maximum observed taint would be. The scale of the vertical axis is not comparable across contests. Within each district, the height of the bar is proportional to the expected workload the four methods under the two error scenarios.

results are shown in an array of barcharts in Figure 4. The y-axes are not comparable across contests: The relative heights of the bars for each district show the relative performance of the methods—shorter bars correspond to a smaller workload. KM (gray) does best in all cases. The three CAST methods are generally comparable, but in some contests (for instance, the last contest on the first row and the second contest on the fourth row) CAST-OPT does noticeably better than CAST-SRS and CAST-PROP.

We found the ratio of the expected workload for the variants of CAST to the workload for KM, since KM had the lowest workload. Figure 5 shows the results. The ratios are between 1.25 and 2.35: KM requires the least hand counting, but some of the CAST methods are almost as efficient for some contests and none is orders of magnitude worse. The three CAST methods are roughly comparable for larger margins, but CAST-OPT has a lower expected workload than the others for smaller margins. Generally, the ratio is higher (KM outperforms CAST more) for smaller margins. This is consistent with the behavior of KM and CAST-SRS in the unstratified case for all 44 contests.

5 Reducing Batch Sizes

In the comparisons above, KM required auditing fewest ballots on average. Workload for all methods depends on batch sizes. Reducing batch sizes generally reduces the workload for risk-limiting audits (at least, when the

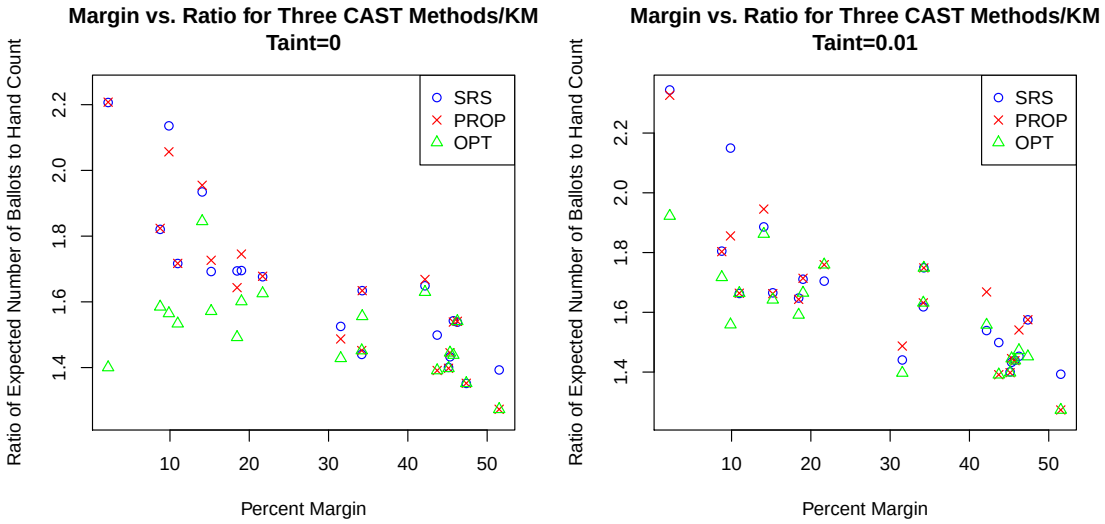

Figure 5: Ratio of expected number of ballots to be audited for three variants of CAST (SRS, PROP and OPT) to the expected number required by KM, plotted against the reported margin for 20 multi-jurisdictional contests at 10% risk limit, for two hypothetical error scenarios. Left panel: the audit finds no errors. Right panel: the audit finds taint 0.01 in every batch.

electoral outcome is correct). In the comparisons above, batches were precincts—precincts were the smallest batches for which the vote-tabulation systems reported vote subtotals. Theory suggests that workload is roughly proportional to the average size of the batches when the electoral outcome is correct [Stark, 2010a]. Doubling the number of batches by cutting every batch in half will halve the workload approximately. Section 6 discusses methods for producing smaller batches in real elections, given the constraints of various commercial vote-tabulation systems and other logistical considerations. This section uses simulations to study the effect of using smaller batches on workload.

5.1 Reducing batch size

We simulated the reduction in workload that accompanies decreasing batch sizes for KM, CAST-SRS, CAST-PROP, and CAST-OPT using the 2008 California House of Representatives data. We use two methods to reduce batch size. The first, "fixed-block reduction," sets a maximum batch size b and splits existing batches so that each new batch has no more than b ballots. Our tests use $b = 50$, in part because this seems large enough to be possible in practice: In November 2009, Marin County, California, divided its ballots into "decks" of no more than 50 ballots. Consider a precinct of size 279. Fixed-block reduction with $b = 50$ divides the ballots from this precinct into six new batches: five of size 50 and one of size 29. Precincts with 50 or fewer ballots are not subdivided. In our tests, we shuffle the ballots in each batch before splitting the batch, so the composition of each new batch is roughly comparable to that of the original precinct. Figure 6 illustrates this procedure for a precinct with 110 ballots.

The second method, "k-cut reduction," divides each precinct into $k \geq 2$ equal sized batches. Precincts of size 0 or 1 are not subdivided. Our tests use $k = 2$, arbitrarily. For example, a precinct of size 200 will be split into two batches of size 100, a precinct of size 2 will be split into two batches of size 1, and a precinct of size 7 will be split into a batch of size 3 and a batch of size 4. As in our tests of fixed-block reduction, we shuffle the ballots in each batch before subdividing the batch.

Fixed-block reduction allows smaller batches to be created as ballots are cast or counted, for instance, by dividing vote-by-mail ballots into "decks" of no more than b ballots. In contrast, k-cut reduction seems

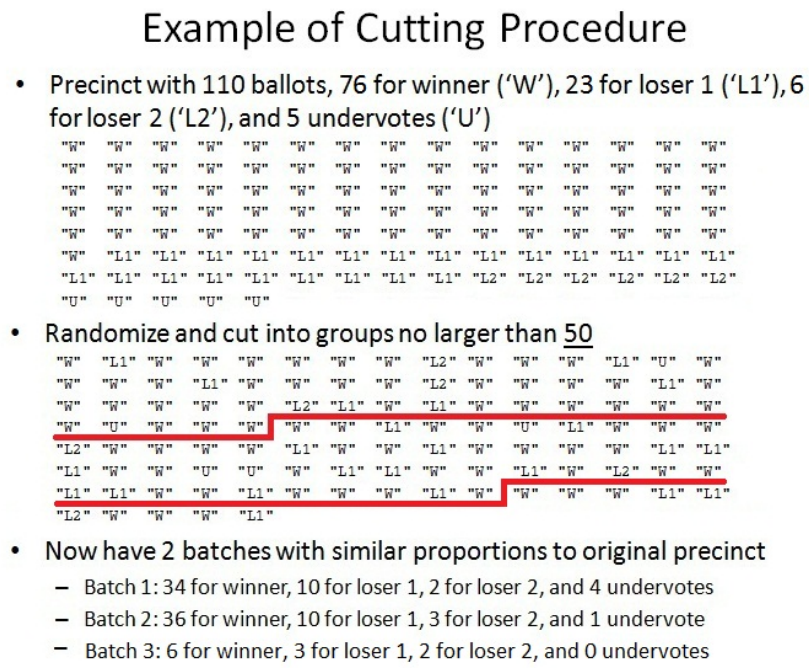

Figure 6: Example of "fixed-block reduction" to divide large batches of ballots into batches with no more than 50 ballots.

to require either knowing how many ballots were cast in each precinct, or taking steps in advance, such as creating k different ballot styles for each precinct [Stark, 2009d]. However, k-cut reduction allows election officials to determine the number of batches ahead of time, while for fixed-block reduction, the number of batches depends on the number of ballots cast in each precinct. Fixed-block reduction gives batches of the same maximum size in different precincts, while k-cut reduction gives the same number of batches for each precinct, of different sizes in different precincts.

We used these two methods to reduce batch size, then simulated the workload for each of the four auditing strategies in each district.

5.2 Defining "typical" batch size

To compare the methods, we introduce summary measures of batch size to quantify the amount by which the methods reduce batch sizes. Recall that N is the number of precincts. Let x_i be the number of ballots cast in precinct i, $i \in \{1, 2, \ldots, N\}$. Let M_i be the number of smaller batches the procedure makes from precinct i, and $M \equiv \sum_{i=1}^{N} M_i$ be the total number of smaller batches. Let x_{ij} be the number of ballots cast in reduced batch j of precinct i, where $j \in \{1, 2, \ldots, M_i\}$.

For k-cut reduction, $M_i = k$ and $x_{ij} \approx x_i/k$: The new batches are clearly k times smaller than they were originally. For fixed-block reduction, things are murkier. Consider the mean across precincts of the ratio of the original precinct size to the average reduced batch size for that precinct:

$$\frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i}{\frac{1}{M_i} \sum_{j=1}^{M_i} x_{ij}} \right) = \frac{M}{N},$$

since $\sum_{j=1}^{M_i} x_{ij} = x_i$. For k-cut reduction, $M/N = k$; for fixed-block reduction, M/N still measures the average reduction in batch size. Figure 7 plots the ratio of the expected workload for the original batches to

the expected workload for the smaller batches against M/N, for fixed-block reduction. (For k-cut reduction, these plots would nearly be straight lines through the origin with slope 1; see table 1.)

We also derive approximate relationships between batch size reduction and workload reduction using only summary statistics from the election: the number of precincts, the number of votes cast for each candidate in each precinct, and the total number of votes cast per precinct. They provide inexpensive estimates of the amount by which reducing batch sizes will reduce workload. We use the ratio

$$\frac{\mathbb{E}\text{ (ballots to audit using precincts as batches)}}{\mathbb{E}\text{(ballots to audit using smaller batches)}}$$

to quantify the reduction in workload.

The reduction in workload is easiest to approximate for KM. Let d_{orig} denote the original number of independent draws for KM and let d_{reduc} denote the number of draws when the batch sizes are reduced. (Recall that KM draws are independent, so the same batch might be drawn more than once; the number of batches drawn is no larger than the number of draws, but can be smaller.)

$$\frac{\mathbb{E}\text{B}_{\text{orig}}}{\mathbb{E}\text{B}_{\text{reduc}}} = \frac{\sum_{i=1}^{N} x_i \left[1 - \left(1 - \frac{u_i}{U_{\text{orig}}}\right)^{d_{\text{orig}}}\right]}{\sum_{i=1}^{N}\sum_{j=1}^{M_i} x_{ij}\left[1 - \left(1 - \frac{u_{ij}}{U_{\text{reduc}}}\right)^{d_{\text{reduc}}}\right]} \approx \frac{\sum_{i=1}^{N} x_i \frac{u_i}{U_{\text{orig}}} d_{\text{orig}}}{\sum_{i=1}^{N}\sum_{j=1}^{M_i} x_{ij}\frac{u_{ij}}{U_{\text{reduc}}} d_{\text{reduc}}}.$$

The approximation holds if for all i and j, $u_i \ll U_{\text{orig}}$ and $u_{ij} \ll U_{\text{reduc}}$.[6] Suppose that $\sum_{j=1}^{M_i} u_{ij} \approx u_i$,[7] so that $U_{\text{orig}} = U_{\text{reduc}}$ and $d_{\text{orig}} = d_{\text{reduc}}$. Then,

$$\frac{\mathbb{E}\text{B}_{\text{orig}}}{\mathbb{E}\text{B}_{\text{reduc}}} \approx \frac{\sum_{i=1}^{N} x_i u_i}{\sum_{i=1}^{N}\sum_{j=1}^{M_i} x_{ij} u_{ij}}.$$

In the next section, we compare this approximation to simulation results.

A similar estimate can be made for CAST-SRS. The approximation is better for k-cut reduction than for fixed-block reduction, as we shall see. The ratio of the original expected workload to the expected workload with smaller batches is

$$\frac{\mathbb{E}\text{B}_{\text{orig}}}{\mathbb{E}\text{B}_{\text{reduc}}} = \frac{\frac{d_{\text{orig}}}{N}\sum_{i=1}^{N} x_i}{\frac{d_{\text{reduc}}}{m}\sum_{i=1}^{N}\sum_{j=1}^{M_i} x_{ij}},$$

where d, N, x_i, M, M_i, and x_{ij} are defined as above. Because $\sum_{i=1}^{N} x_i = \sum_{i=1}^{N}\sum_{j=1}^{M_i} x_{ij}$ (i.e., the total number of ballots cast in the district is the same before and after the reduction procedure), this reduces to

$$\frac{\mathbb{E}\text{B}_{\text{orig}}}{\mathbb{E}\text{B}_{\text{reduc}}} \approx \frac{d_{\text{orig}} m}{d_{\text{reduc}} n}. \tag{1}$$

To calculate (1) requires knowing d_{orig} and d_{reduc}. We approximate them as follows: For any given contest, let α be the risk limit and let n be the total number of batches. Suppose that at least q of the n batches must have taint greater than t if the apparent outcome is incorrect. Find

$$d_0 \equiv \min\left\{ d : \frac{\binom{N-q}{d}\binom{q}{0}}{\binom{N}{d}} \le \alpha \right\},$$

the number of draws necessary to certify the apparent outcome if no batches are observed to have taint greater than t. We assume that $d \ll n$, so sampling without replacement is almost equivalent to sampling with replacement[8]. Hence,

[6] Across all 44 districts, the largest value of the maximum u_i in the district divided by U_{orig} is 9.77×10^{-3}; the largest value of the maximum of u_{ij} in the district divided by U_{reduc} is 5.19×10^{-4}.

[7] Across all 44 districts, the maximum value of $|u_i - \sum_{j=1}^{M_i} u_{ij}|$ is 5.55×10^{-17}.

[8] This is not necessarily the case. For example, in District 4 the margin in votes was less than the number of votes in a single precinct, so $d \approx n$. Figure 7 shows that the approximation is not accurate in that case.

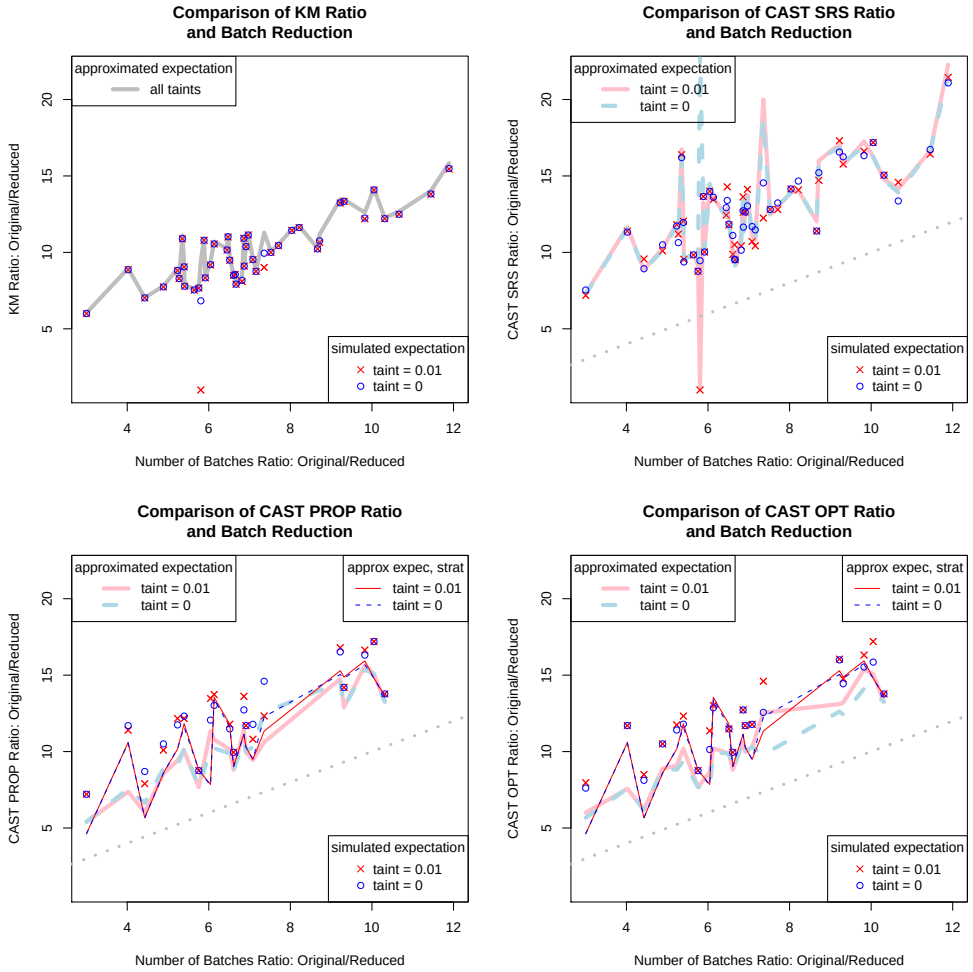

Figure 7: Ratio of the expected workload using precincts as batches to the expected workload for smaller batches, plotted against the average reduction in batch size m/n, for (top left) KM, (top right) CAST-SRS, (bottom left) CAST-PROP, and (bottom right) CAST-OPT. Each point represents a district. The plots for KM and CAST-SRS show 44 districts; the plots for CAST-PROP and CAST-OPT show only the 20 districts that cross county lines. The thick lines show a simple approximation to the ratio of the expected workload that does not take stratification into account. The dotted line in the three CAST plots is the forty-five degree line. The thin lines in the CAST-PROP and CAST-OPT plots show a simple approximation for the stratified case.

$$
\begin{aligned}
d_0 &\approx \min\left\{d : \left(\frac{N-q}{N}\right)^d \leq \alpha\right\} \\
&= \min\left\{d : d\ln\left(\frac{N-q}{N}\right) \leq \ln\alpha\right\}. \qquad (2)
\end{aligned}
$$

If $q \ll n$ (that is, if the number of tainted batches necessary to alter the outcome is much smaller than the total number of batches), then $d\ln\left(\frac{N-q}{N}\right) \approx -d\frac{q}{N}$ and thus

$$
d_0 \approx \frac{-(\ln\alpha)\,n}{q}. \qquad (3)
$$

Substituting (3) into (1) yields

$$
\frac{\mathbb{E}\mathrm{B}_{\mathrm{orig}}}{\mathbb{E}\mathrm{B}_{\mathrm{reduc}}} \approx \frac{\frac{Nm}{q_{\mathrm{orig}}}}{\frac{mn}{q_{\mathrm{reduc}}}} = \frac{q_{\mathrm{reduc}}}{q_{\mathrm{orig}}}. \qquad (4)
$$

For an even simpler approximation, we could consider q to be proportional to the number of batches so that $q_{\mathrm{reduc}}/q_{\mathrm{orig}} \approx m/n$. Using this approximation in (3) gives $d_{\mathrm{orig}} \approx d_{\mathrm{reduc}}$, so

$$
\frac{\mathbb{E}\mathrm{B}_{\mathrm{orig}}}{\mathbb{E}\mathrm{B}_{\mathrm{reduc}}} \approx m/n. \qquad (5)
$$

This is shown in Figure 7 as the gray dotted line with slope 1. The accuracy of this approximation depends on how q varies with batch size. It should be better when the original batch sizes are roughly equal than when they vary widely, and better for k-cut reduction than for fixed-block reduction. If the original batch sizes are uneven, k-cut reduction will not increase q in direct proportion to k.

Fixed-block reduction generally does not increase q in proportion to the ratio of new batches to old batches. As a result, the approximation (5) will tend to be poor for fixed-block reduction, as confirmed by the fact that in Figure 7, the ratios are closer to the approximate relationship than they are to the forty-five degree line.

For CAST-SRS, the outcome-changing errors that are hardest to detect are those that concentrate in the smallest number of batches: q controls the amount of sampling required to confirm the outcome. For stratified sampling, that is not the case: the outcome-changing error that is hardest to detect depends in detail on the distribution of error bounds within each stratum. This makes it hard to develop an approximation to the reduction in workload for stratified sampling that is both simple and accurate. We develop a simple but crude approximation based on q; better approximations would use more information about the number of batches that must have errors in the outcome-changing scenario that is hardest to discover using the particular stratified sampling design. A calculation including this information would be almost as computationally intensive as finding the true expected value, so we use a more crude approximation. Because our approximation depends only on q, it is the same for CAST-PROP and CAST-OPT.

Recall that N_c is the number of batches of original size in stratum c. Let M_c be the number of batches of reduced size in stratum c, $q_{c,\mathrm{orig}}$ be the number of original batches in stratum c that must have error for the apparent outcome to be incorrect, for the distribution of error that minimizes the chance that the sample contains any taint greater than t. Let $q_{c,\mathrm{reduc}}$ be the corresponding number after batches have been divided. Then the approximation becomes

$$
\frac{\mathbb{E}\mathrm{B}_{\mathrm{orig}}}{\mathbb{E}\mathrm{B}_{\mathrm{reduc}}} \approx \frac{\sum_{c=1}^{C} \frac{N_c M_c}{q_{c,\mathrm{orig}}}}{\sum_{c=1}^{C} \frac{M_c n_c}{q_{c,\mathrm{reduc}}}}. \qquad (6)
$$

These approximations are plotted as thin lines in the two bottom panels of Figure 7. Note that in Figure 7 the approximations for CAST-PROP and CAST-OPT are not as accurate as the approximations for KM and CAST-SRS. Nevertheless, the approximations for the two stratified cases provide ballpark estimates of workload savings.

Method	Taint	Contests	Min	Q1	Median	Mean	Q3	Max
KM	0	43	1.8709	1.9845	1.9906	1.9854	1.9938	1.9972
	0.01	43	1.7804	1.9834	1.9906	1.9822	1.9938	1.9972
CAST-SRS	0	43	1.5598	1.9542	1.9812	1.9628	1.9916	2.0769
	0.01	43	1.5623	1.9545	1.9817	1.9621	1.9915	2.0751
CAST-PROP	0	20	1.5706	1.9564	1.9890	1.9514	1.9966	2.0400
	0.01	20	1.5670	1.9548	1.9815	1.9448	1.9915	2.0382
CAST-OPT	0	20	1.7597	1.9537	1.9834	1.9681	1.9913	2.0618
	0.01	20	1.5706	1.9524	1.9850	1.9546	1.9947	2.0400

Table 1: Workload ratio for "k-cut reduction" with k=2. Summary statistics for the ratio of the expected number of ballots to audit using batches of original size to the expected number of ballots to audit using batches of reduced size, for KM, CAST-SRS, CAST-PROP, and CAST-OPT, for two error scenarios: taint t is zero for all audited batches, and taint $t = 0.01$ for all audited batches. For each contest, k-cut reduction was repeated 50 times using different random orderings of the ballots. District 4 is omitted. Column 1: auditing method. Column 2: taint of all observed batches. Column 3: number of contests. Columns 4–9: minimum, lower quartile, median, mean, upper quartile, and maximum of the ratio across the contests. The ratios concentrate near 2, as expected.

5.3 Results for k-Cut Reduction

As discussed above, it is easier to anticipate the reduction in workload for k-cut reduction because $M_i = k$ and $x_{ij} \approx x_i/k$. For $k = 2$, the expected number of ballots to audit to certify the apparent outcome after reduction is roughly half of the number of ballots required for the original batches. Table 1 summarizes the ratio $\mathbb{E}\mathrm{B}_{\mathrm{orig}}/\mathbb{E}\mathrm{B}_{\mathrm{reduc}}$ for k-cut reduction. The ratios concentrate near 2, as predicted.

6 Reducing batch sizes in practice

We have demonstrated that reducing batch sizes can drastically decrease the number of ballots that need to be audited to certify the apparent outcome of the election. However, current federally certified vote-tabulation systems (VTSs) generally do not support reporting by batches smaller than precincts, perhaps separately for votes cast by mail and votes cast in person. Such batches vary in size from a handful of ballots to about two thousand ballots. In this section, we discuss some methods for reducing batch size in practice that could be used with current VTSs, and the costs and limitations of those methods.

To base an audit on batches smaller than precincts, two things must happen in practice. First, the VTS must be able to generate and export subtotals for those batches. Second, the physical ballots (or the audit trail, such as a voter-verifiable paper audit trail or VVPAT) that correspond to each subtotal must be identifiable and retrievable for a hand-to-eye count if that batch is selected for audit.

Ideally, the VTS would report the interpretation of individual ballots (cast vote records) and allow auditors to associate individual cast vote records with the corresponding physical ballot. This would make ballot-level audits (audits using batches consisting of single ballots) possible.[9] The next generation of systems promises to export cast vote records; for most current VTSs, ballot-level auditing requires using a "shadow" system in addition to the system of record. (The pilot audits under California AB 2023 will include ballot-level audits using shadow systems.[10])

[9] For a proposed methodology for risk-limiting ballot-level audits, see Stark [2010b]; Benaloh et al. [2011]. For a discussion of the statistical advantages of auditing at the ballot level, see Neff [2003]; Stark [2010a].

[10] Audits at the ballot level have been conducted in Yolo County, CA, in November 2009 [Stark, 2009b] and in Orange County, CA, in March 2011 (Stark, manuscript in preparation). An audit based on "ballot polling" was conducted in Monterey County, CA, in May 2011 (Stark, manuscript in preparation). Ballot polling involves drawing a random sample of paper ballots and interpreting them, but not comparing them to their interpretation according to the voting system. Ballot polling can be used for risk-limiting audits by examining ballots at random until one can reject the hypothesis that the apparent winner received a smaller fraction of the votes than any of the other candidates.

Reporting subtotals for extremely small batches might reduce voter privacy and introduce greater opportunity for vote selling or voter coercion. However, if voters cannot be matched to batches (for instance, if batches are formed by randomly subdividing the ballots cast in precincts) and if contests are reported separately (so a voter cannot signal the identity of his ballot in the reported results by voting in a pattern), these risks seem small—certainly no larger than they are currently for small precincts. In any event, results could be publicly reported for larger batches than the audit relies on, for instance, precincts. The election officials need to "commit" to the auditable subtotals before the audit starts, but that could be accomplished by putting the data in escrow rather than by publishing the data for very small batches. For an alternative approach, see Benaloh et al. [2011].

How can we reduce batch sizes? Reporting votes cast in-person separately from early voting and vote by mail helps. For direct recording electronic voting machines (DREs; i.e. touch screen voting machines) with VVPATs, we could audit machine results instead of precinct results. Some VTSs can track results separately by machine. If the VVPATs corresponding to a given machine can be identified, machine-level subtotals can be audited against the paper trail; this approach was used in a risk-limiting audit in Orange County, CA, in March 2011, where pollworkers were instructed to distribute voters evenly among the DREs in each precinct, which reduced batch sizes by a factor of about 10 [Stark, 2011a].

Jurisdictions that use central-count optical scan systems (CCOS) could reduce batch sizes by grouping ballots into small "decks" of no more than b ballots before counting, provided the VTS can report subtotals by deck and the jurisdiction tracks and stores decks in a way that allows the deck corresponding to a reported subtotal to be retrieved and counted by hand if the batch is selected for audit. This approach was tested in Marin County, CA, in November 2009 [Stark, 2009b].

For jurisdictions that use precinct-count optical scan systems (PCOS), batch sizes could be reduced without increasing the burden on pollworkers in the following way. VTSs can track votes by ballot style. A "ballot style" is the specific collection and ordering of contests and candidates on a ballot. Typically, the ballots in a single precinct are all of one ballot style, or of a small number of ballot styles if only some voters in the precinct are eligible to vote in one or more of the contests in the election (this is common in primary elections, where ballot styles often depend on party affiliation). Artificially increasing the number of ballot styles by marking groups of no more than 100 ballots with a barcode to identify them as a batch would allow existing software to track and report subtotals for those batches. It would not be necessary for jurisdictions to account for each ballot pseudo-style sent to a precinct separately; the difference between the styles is solely to enable the VTS to tally subtotals for each batch and so that—if the batch is selected for audit, the ballots that comprise that batch can be identified.

Using pseudo-ballot-styles would increase printing costs modestly, but would greatly reduce audit costs. On the other hand, it could greatly increase the costs of logic and accuracy testing, if every ballot pseudo-style needs to be tested individually. But perhaps logic and accuracy testing could be done with a mix of ballot pseudo-styles for each precinct (say, half a dozen of each pseudo-style, mixed together), instead of testing each ballot pseudo-style separately.

Here is a sketch of how one might use pseudo-styles. Consider a precinct with 800 registered voters, of whom 300 request absentee ballots. The jurisdiction would print three batches of 100 vote by mail ballots that are identical except for a barcode and a letter (A–C). The jurisdiction would print (up to) five batches of ballots to be used in the precinct, identical except for a barcode and a letter (D–H). When the ballots are scanned—whether using CCOS or PCOS—the barcode would make it possible for the VTS to subtotal batches of no more than 100 ballots. Changes may be required to make it possible for the VTS to export those subtotals in a useful format. And those changes might require new federal and state certification.

If the audit selects one of the batches from a precinct to count by hand, the ballots in that precinct would then be sorted manually (using the letter code) or with an automated sorter (using the barcode) to isolate the batch that is to be counted by hand. There would be no need to sort or separate batches in precincts in which no batches are selected to be audited.

Using small batches raises privacy concerns. Steps should be taken to ensure that pollworkers do not use the proliferation of ballot pseudo-styles to determine how particular subgroups of voters voted, for instance, by giving all senior citizens ballots of one style. Omitting the human-readable letter would improve voter

privacy, since—on the assumption that people do not routinely read barcode—neither the voter, pollworkers, nor elections officials would know which pseudo-ballot-style a given voter received. To increase voter privacy, the ballot pseudo-styles could be shuffled before handing ballots to voters, so that there is no way to know which batch contains a given voter's ballot.

Reducing batch sizes can greatly reduce the burden of hand counting, as our simulations confirm. It could increase costs in other ways, though. For instance, the physical batches of ballots need to be retrievable, which entails some organizational costs. Using ballot pseudo-styles would increase printing costs.[11] Reducing the size of the "remainders" could reduce some waste: For instance, if a precinct has 515 registered voters, one might print 75-ballot batches, so that the remainder is $515 - 450 = 65$ (wasting $75 - 65 = 10$ ballots), rather than $515 - 500 = 15$ (wasting $50 - 15 = 35$ ballots).

There is clearly room for clever solutions, although efficient approaches might require changes to laws and regulations. Longer term, voting systems should be designed to facilitate auditing by creating and reporting subtotals for small, physically retrievable batches.

7 Conclusions

Risk-limiting post election audits improve on previous election auditing methods by guaranteeing a large chance of correcting incorrect outcomes. CAST (Canvass Audits by Sampling and Testing) and KM (Kaplan-Markov) are approaches to risk-limiting audits. Both rely on cluster samples: samples of batches of ballots for which the voting system is able to report results. CAST-SRS is a variant of CAST that uses simple random sampling of batches. KM uses a sample drawn with probability proportional to a bound on the error in each batch; the draws are independent. Comparisons using data from the 2008 U.S. House of Representatives contests in California (and hypothetical errors) suggest that KM almost always requires less work than CAST-SRS. Gains are largest when the margin is small.

In cross-jurisdictional elections, stratifying the sample by jurisdiction may be necessary for logistical reasons. We study two variants of CAST that use stratified sampling. Both take simple random samples within strata, and sample independently across strata. CAST-PROP takes the sample size in each stratum to be proportional to the number of batches in the stratum. CAST-OPT optimizes the sample size in each stratum to minimize the number of batches to audit, on the assumption that observed errors will be less than a pre-specified threshold. The expected workload for these stratified methods is larger than for KM, but KM requires an unstratified sample.

Reducing batch size can reduce audit workload when the contest outcome is correct. We consider two generic approaches to reducing batch sizes: subdividing batches into smaller batches with a given maximum size, and subdividing batches into k smaller batches of (essentially) equal size. Simulations show that the reduction in workload is roughly proportional to the reduction in batch size, as theory predicts. Auditing workload could be reduced by several orders of magnitude by using individual ballots as batches rather than using precincts as batches. Auditing using individual ballots as batches may require changes to current voting systems or the use of "shadow" systems, but they are extremely efficient and risk-limiting rules for ballot level audits can be quite simple [Stark, 2010b].

Many groups concerned with election integrity endorse risk-limiting audits as best practice. Pilot audits, theoretical work, and the results we present here show that risk-limiting audits can be performed very economically—if batch sizes can be reduced substantially. "Data plumbing" to enable voting systems to report results in a useful machine-readable format, for small batches of ballots that can be located and counted by hand, is the key to efficient risk-limiting audits.

[11] The increase depends on many things, including whether the jurisdiction delivers the printer camera-ready .pdf of the ballots.

Acknowledgments

We are grateful to Mike Higgins, Luke Miratrix, and Hua Yang for the use of software they wrote and for helpful conversations. We are grateful to Jennie Bretschneider and Mark Lindeman for helpful conversations.

References

Aslam, J., Popa, R., and Rivest, R. (2007). On auditing elections when precincts have different sizes. people.csail.mit.edu/rivest/AslamPopaRivest-OnAuditingElectionsWhenPrecinctsHaveDifferentSizes.pdf.

Benaloh, J., Jones, D., Lazarus, E., Lindeman, M., and Stark, P. (2011). Soba: Secrecy-preserving observable ballot-level audits. Technical report, Dept. Statistics, Univ. of Calif., Berkeley.

Hall, J. L., Miratrix, L. W., Stark, P. B., 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.

Higgins, M., Rivest, R., and Stark, P. (2011). A sharper treatment of stratification in risk-limiting post-election audits. *Working draft.*

Institute of Governmental Studies, University of California, Berkeley (2010). The statewide database. `http://swdb.berkeley.edu/`. The redistricting database for the state of California.

Lindeman, M., Halvorson, M., Smith, P., Garland, L., Addona, V., and McCrea, D. (2008). Principles and best practices for post-election audits. `www.electionaudits.org/files/best%20practices%20final_0.pdf`. Retrieved April 20, 2011.

McBurnett, N. (2010). Boulder County 2010 General Election risk-limiting audit – County Coroner race. `http://bcn.boulder.co.us/~neal/elections/boulder-audit-10-11/`.

Miratrix, L. (2009). Elec: Collection of functions for statistical election audits. `http://cran.r-project.org/web/packages/elec/`.

Neff, C. (2003). Election confidence: A comparison of methodologies and their relative effectiveness at achieving it. Technical report, VoteHere, Inc. Retrieved March 6, 2011, from `http://www.verifiedvoting.org/downloads/20031217.neff.electionconfidence.pdf`.

Saldaña, L. (2010). California Assembly Bill 2023. `www.leginfo.ca.gov/pub/09-10/bill/asm/ab_2001-2050/ab_2023_bill_20100325_amended_asm_v98.html` Retrieved April 20, 2011.

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. (2009d). The status and near future of post-election auditing. `http://statistics.berkeley.edu/~stark/Preprints/auditingPosition09.htm`.

Stark, P. (2010a). Risk-limiting vote-tabulation audits: The importance of cluster size. *Chance*, 23(3):9–12.

Stark, P. (2010b). Super-simple single-ballot risk-limiting audits. 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/Stark.pdf`. Retrieved April 20, 2011.

Stark, P. (2011a). 2011 Pilot risk-limiting audits under California AB 2023. *Working draft*.

Stark, P. (2011b). Simple approaches to risk-limiting audits. *Working draft*.