# Modeling Cost of Risk-Limiting Audits on U.S. Federal Elections

Jonathan Brasch Yuxin Li Aditya Middha Jay Schauer

## Abstract

Risk-limiting audits (RLAs) provide an efficient way to confirm initial elections results with statistical methods. Ballot-polling and ballot-level comparison RLAs are modeled on president and congressional elections between 1976 and 2018. These methods provide an estimated number of ballots which must be drawn to confirm initial election results when a RLA is run with a risk limit of 10%. Cost is estimated based based on previous RLAs run by state governments on federal elections. From this, an upper-bound on cost is estimated for all federal race within the period to be audited individually and some federal races within a state to be audited simultaneously. Based on estimated past costs, future costs are projected. Distribution of auditing work between the states is discussed as well as how these cost estimates compare to current recount costs.

## 1 Introduction

A functioning democracy requires the trust of the citizens in the electoral process. Currently, most jurisdictions in the United States rely on machines to tally votes [17]. Due to this, for a citizen to place trust the process of counting votes they must trust machines to accurately scan and record their paper ballots. However, these computerized systems may produce incorrect results due errors in how machine or humans administer elections or deliberate subversion of results. Given the high stakes of Americans elections as well as adversaries' demonstrated interest in influencing results, it is clear that double-checking the validity of process and result of tallying votes is necessary. This process of double checking is known as a post-election audit.

While twenty eight states finalize official election results without verifying computer-tabulated vote totals, nineteen states in the US have some post-election audit requirements [3]. However, most of those procedures are at some level arbitrary – they sometimes require considerably more work than actually needed to confirm the election result while sometimes require far too little information. RLAs are designed to perform just the right amount of work to confirm the election result. Instead of manually recounting every vote to gain the certainty of the election outcome, the number of votes we need to manually count depends on the margin of victory in the contest that we are auditing [12]. That means generally (in races that are not very competitive) a state-wide RLA of some election may only require checking less than 1,000 to 10,000 paper ballots to validate the election result. Thus, RLAs are an efficient tool to confirm the correct election outcomes.

Since RLA provide an effective and efficient way to confirm election results, Congress may consider requiring the states and D.C. to perform RLAs for federal races to reach a better level of certainty on the election outcome. Three major determining factors in making this decision are the overall workload of performing nationwide RLAs, how this burden is distributed among the states, and the amount of money needed to perform this RLA. We estimate RLA workloads and model audits costs using real result data from the 2012 to 2018 elections and actual costs from the jurisdictions that have implemented RLAs so far.

By estimating the overall workload and modelling the audit costs using data from actual recent elections, we demonstrate the feasibility of requiring the states and D.C. to perform RLAs for all federal races. By having the estimated result of the audits' workload and costs, the states and Congress have more evidence in considering its benefits with acceptable workload and costs.

## 2 Background

The idea of RLAs originates from a 2007 paper [23] by Dr. Philip B. Stark [1], titled "Risk-Limiting Post Election Audits: Conservative P-Values From Common Probability Inequalities", where he details a process in which jurisdictions can provide strong statistical evidence that the tallied winner is the true winner of an election or have a high probability of cor-

[1] Associate Dean, Division of Mathematical and Physical Sciences, University of California, Berkeley

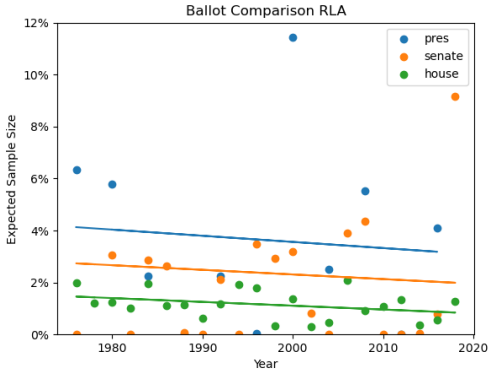


Figure 1: Estimated ballots drawn to fulfill a ballot-level comparison audit with a risk limit of 10% as a proportion of total ballots cast for presidential, senate, and house elections from 1976-2018. Solid lines denote a linear regression for each group of elections.

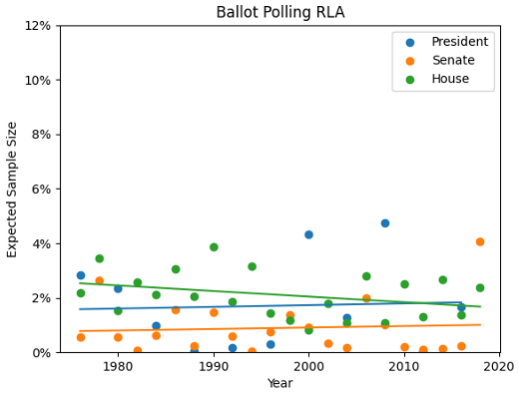


Figure 2: Estimated ballots drawn to fulfill a ballot-polling audit with a risk limit of 10% as a proportion of total ballots cast for presidential, senate, and house elections from 1976-2018. Solid lines denote a linear regression for each group of elections.

recting an incorrect initial count. The aptly named "risk limit" $\alpha$ is a specified value (often between 5-10%) that represents the maximum probability that an audit will not continue to a full hand tally when the original outcome is wrong; in other words, the audit limits the risk of an erroneous result. For instance, an audit that has at least a 95% chance of correcting an incorrect outcome has a 5% risk limit.

All RLAs begin with the requirement of a voter-marked paper record – either marked manually or using ballot marking devices – so that there is a convenient means to access the correct intention of all voters. Additional requirements include a specified chain of custody for these ballots and a successful compliance audit in which ballot security is assured and a ballot manifest (a description of how ballots are organized in containers) is created. From this point on, not all RLAs are uniform in their mechanics and vary significantly to match the technology and resources available [12].

Ballot-polling audits make no demands of a vote tabulation system, as simply knowing the reported winner and the margin will suffice to examine a random sample of individual ballots. With $\alpha = 10\%$ and starting with $T = 1$, two separate multipliers, based on the margin $m$ and the fault tolerance $\lambda$, are used to influence the confidence in the initial outcome. While going through the audit samples, observers know that if the $T$ value reaches lower than 0.011, a full hand count is issued, and if $T$ rises above 9.9, the audit can stop [12].

On the other hand, ballot-level comparison audits require less vote sampling than ballot-polling audits but do rely on vote tabulation systems to compare each sampled ballot to its machine interpretation. These audits rely on four constants: reported margin $m$, risk limit $\alpha$, error inflation factor $\gamma$ and error tolerance factor $\lambda$. $\gamma \geq 100\%$ controls the trade-off between sample size and additional counting with overstatements, and $\lambda < 100\%$ acts as a fixed penalty for overstatements. Election administrators can decide $\alpha$, $\gamma$, and $\lambda$ before the election. After an initial sample of the total ballots, additional ballots may be drawn depending on the number of understatements (where fixing the error causes an increase in the winner's margin) and overstatements (where fixing the error causes an decrease in the winner's margin) in the sample. Depending on these factors, this audit method will either halt sampling or continue to a full hand recount [12, 24]. Unfortunately, very few federally certified vote tabulation systems provide machine interpretations of individual ballots creating a cast vote record (CVR), but these improvements are likely to arrive with the next generation of vote tabulation systems. One way around this is the process of transitive auditing, which usually entails re-tabulation through third-party scanners. Batch comparison audits, in which entire batches are counted manually, offer an alternative that does not require vote tabulation technology, but typically require examining the largest number of ballots [3].

It is often the case that multiple races fall under one election authority. Here, it may be more efficient to run a simultaneous RLA, in which all contests are audited at once, than multiple separate audits. In a simultaneous RLA, the initial sample size depends on the diluted margin $\mu$, the smallest margin by percentage in votes across all contests. Similar to a ballot-level comparison audit of a single race, a simultaneous ballot-level comparison audit relies on an error inflation factor $\gamma$ and error tolerance factor $\lambda$. These constants, along with $\mu$, determine the initial sample size [24]. Similar to an individual race audit, the simultaneous audit relies on the number of overstatements (normalized across all races) as its halting criteria. Overall, the main benefit of conducting a simultaneous audit is that separate samples are no longer needed to audit multiple races, extremely helpful in decreasing the work and time requirements in election years.

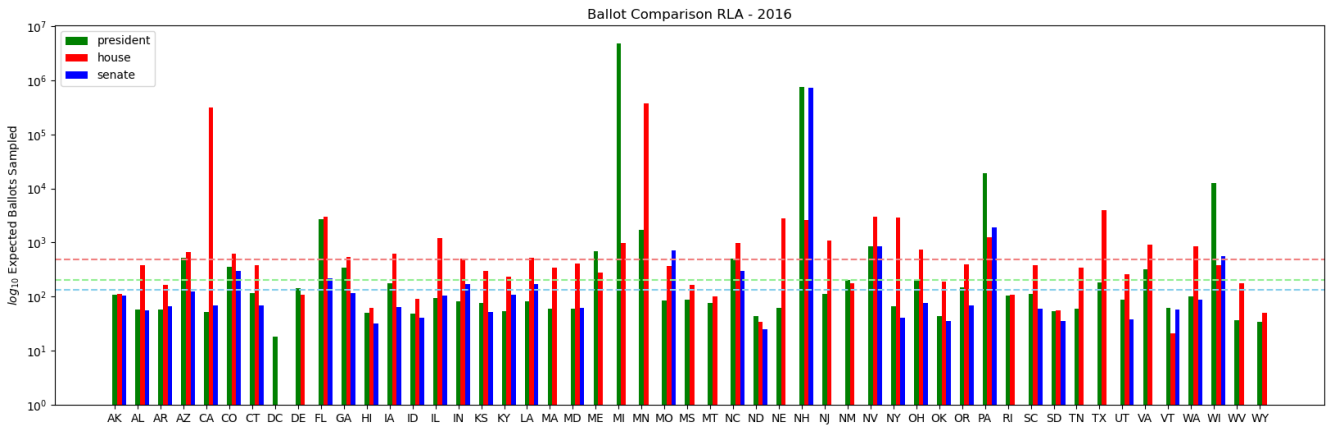


Figure 3: Estimated number of ballots drawn to fulfill a ballot-level comparison audit with a risk limit of 10% per state for presidential, senate, and house elections from 1976-2018. Dashed lines denote average number of ballots drawn for each group of elections.

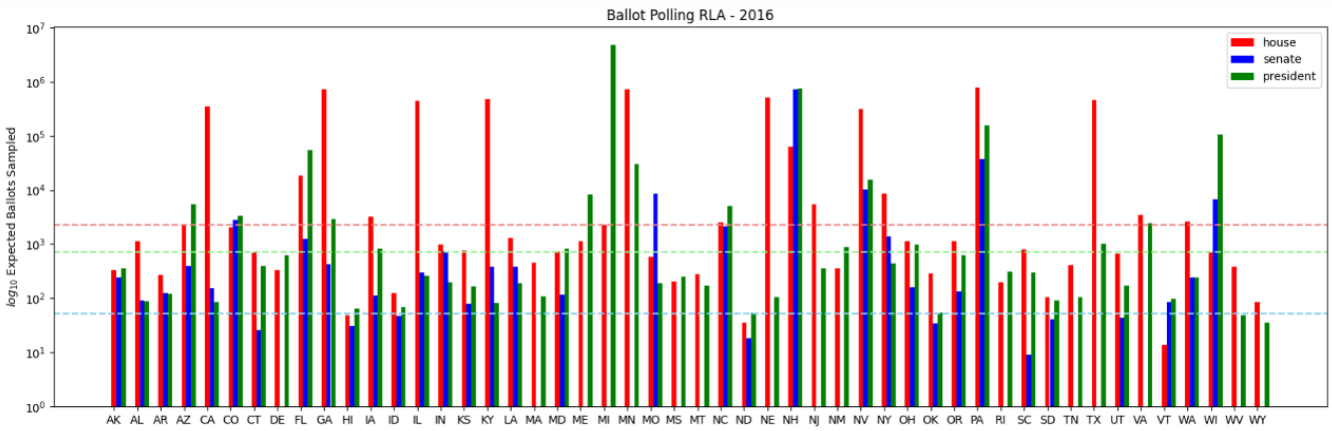


Figure 4: Estimated number of ballots drawn to fulfill a ballot-polling audit with a risk limit of 10% per state for presidential, senate, and house elections from 1976-2018. Dashed lines denote average number of ballots drawn for each group of elections.

Another consideration for modeling RLAs for federal elections is their plurality nature [2] as the regular ballot-polling RLA schema hinges on a majoritarian system. In this alternate setup, the auditors first set value $M$ to be a soft limit for the maximum number of ballots to audit before requiring a full hand count and value $s_w$ to be the reported proportion of votes cast for candidate $w$. As the audit proceeds, if the ballot shows a vote for $w$, test statistic $T$ is multiplied by $2s_w$, and is otherwise multiplied by $2(1-s_w)$. As a result, there are now two stopping criteria, (1) if $T$ is greater than or equal to 1 divided by the risk limit $\alpha$ or (2) if $M$ ballots have been counted [13].

[2] While most states election their congressional delegation using a first-past-the-post plurality method, some (such as GA, LE, ME) require an outright majority to win. In the event no candidate reaches this threshold, the races advances to a runoff. This could be instant to a second special election. A candidate for president requires a majority of the electoral college, but only needs a plurality to win a state's electoral votes

Altogether, one can understand the purpose of the RLA to be twofold: not only does it prevent wasting resources on full hand recounts by halting recounts earlier given sufficient evidence, it also helps correct machine error with comparison to a verifiable audit trail (for ballot-comparison audits). These two issues are more commonplace than the average citizen may think, with just the 2020 election cycle bringing a week-long full hand recount for the presidential election in Georgia and election software issues in Antiram County, Michigan [8, 15]. Additional examples include a two-week long recount in the 2016 Wisconsin presidential election, and significant voting machine errors in the 2008 New Jersey primary, 2004 Ohio presidential, and 2004 California primary races. Most states throughout the country do have some form of a recount procedure, but can be seen as arbitrary (for example, each California county recounts the votes in 1% of precincts) compared to the statistics-based evidence that RLAs provide [22].

| Year | House | Senate | President | President & Senate | Total |
|---|---|---|---|---|---|
| 1976 | 1,475,364 | 5,031 | 5,174,928 | 5,175,782 | 6,651,146 |
| 1978 | 661,976 | 3,544,105 | | 3,544,105 | 4,206,081 |
| 1980 | 977,190 | 1,817,607 | 4,999,906 | 6,815,125 | 7,792,315 |
| 1982 | 652,994 | 8,278 | | 8,278 | 661,272 |
| 1984 | 1,605,620 | 1,297,044 | 2,089,961 | 3,384,640 | 4,990,260 |
| 1986 | 671,620 | 1,239,338 | | 1,239,338 | 1,910,958 |
| 1988 | 933,932 | 47,973 | 12,556 | 57,070 | 991,002 |
| 1990 | 397,368 | 4,159 | | 4,159 | 401,527 |
| 1992 | 1,147,250 | 1,508,563 | 2,349,809 | 3,852,616 | 4,999,866 |
| 1994 | 1,360,504 | 6,451 | | 6,451 | 1,366,955 |
| 1996 | 1,587,104 | 1,708,698 | 24,534 | 1,728,817 | 3,315,921 |
| 1998 | 211,447 | 1,585,660 | | 1,585,660 | 1,797,107 |
| 2000 | 1,364,128 | 2,468,899 | 12,058,117 | 14,523,778 | 15,887,906 |
| 2002 | 227,806 | 343,165 | | 343,165 | 570,971 |
| 2004 | 503,595 | 11,900 | 3,083,229 | 3,092,289 | 3,595,884 |
| 2006 | 1,675,725 | 2,380,931 | | 2,380,931 | 4,056,656 |
| 2008 | 1,139,602 | 2,894,114 | 7,246,950 | 10,138,077 | 11,277,679 |
| 2010 | 944,107 | 6,148 | | 6,148 | 950,255 |
| 2012 | 1,636,555 | 13,121 | 14,415 | 23,896 | 1,660,451 |
| 2014 | 277,372 | 16,413 | | 16,413 | 293,785 |
| 2016 | 723,128 | 745,945 | 5,586,512 | 5,589,662 | 6,312,790 |
| 2018 | 1,444,407 | 8,196,437 | | 8,196,437 | 9,640,844 |

Table 1: Estimated total ballots drawn to fulfill a ballot-level comparison audit with a risk limit of 10% for all house elections, senate elections, presidential elections in all states and D.C, elections for president and senate simultaneously, and for all federal races per year. In cases where the sum of ballots expected to be audited for senate and president was less than a modeled simultaneous audit for a given state, the better case was consider for the President-Senate results as well as to the national total.

Compounding the theoretical promise of RLAs, several pilot RLAs have been conducted over the past decade to prove the effectiveness and efficiency of the process. The first reported RLA was conducted by the Stark in 2008, soon after his paper establishing the process. In this, Stark's team found great success in reducing the sample size (frequently audited less than 1% of ballots cast), but also observed bottlenecks in pre-processing ballots to a usable format, adapting to the reduced technological expertise amongst election staff, and conducting simultaneous audits. The question of simultaneous audits was solved with new probability formulations in a 2009 local election RLA in Yolo, CA; it would not be until a 2011 pilot in California where Stark's team was able to vastly improve efficiency of RLAs by introducing web-based audit tools that automate calculations and randomized selection of ballots [9]. A follow-up 2013 pilot by Stark's team reduced the pre-processing effort by utilizing commercial off-the-shelf (COTS) scanners to individually pair each ballot with a scanned ballot image, but also suggested that voting systems become capable of producing CVRs for each ballot to further increase efficiency. This last Stark study appears to be a turning point, as several states have gone on to conduct their own pilot RLAs, armed with the same web-based audit tools and COTS scanners, and noticed significant reductions in cost, logistical planning, and workload [19]. Cost and timing statistics from these studies were used to inform this paper's modeling of nationwide RLA costs. With RLAs becoming more of a proven option, an increasing number of states have established a requirement for RLAs thus far: Colorado, Georgia, Indiana, Nevada, New Jersey, Rhode Island, Virginia, and Michigan are conducting RLAs or implementing a pilot program for the 2020 election [16].

## 3 Data & Methods

### 3.1 Election Results

We use official election results from federal races between 1976 and 2018 [4–6]. These data report elections for the president and senate by state and house elections by congressional district. In our analysis, we only consider general elections for voting members of Congress; this elides primaries, runoffs, special elections, and elections for congressional delegates from territories and the District of Columbia.

Moreover, we make two further simplifications. First, we consider all elections to be held with a first-past-the-post plurality method. A small number of states require an outright majority to win a congressional seat in a general election;

| Year | House Cost | Senate Cost | President Cost | President & Senate Cost | Total Cost |
|---|---|---|---|---|---|
| 1976 | $7,317,805 | $24,953 | $25,667,642 | $25,671,878 | $32,989,684 |
| 1978 | $3,283,400 | $17,578,760 | | $17,578,760 | $20,862,161 |
| 1980 | $4,846,862 | $9,015,330 | $24,799,533 | $33,803,020 | $38,649,882 |
| 1982 | $3,238,850 | $41,058 | | $41,058 | $3,279,909 |
| 1984 | $7,963,875 | $6,433,338 | $10,366,206 | $16,787,814 | $24,751,689 |
| 1986 | $3,331,235 | $6,147,116 | | $6,147,116 | $9,478,351 |
| 1988 | $4,632,302 | $237,946 | $62,277 | $283,067 | $4,915,369 |
| 1990 | $1,970,945 | $20,628 | | $20,628 | $1,991,573 |
| 1992 | $5,690,360 | $7,482,472 | $11,655,052 | $19,108,975 | $24,799,335 |
| 1994 | $6,748,099 | $31,996 | | $31,996 | $6,780,096 |
| 1996 | $7,872,035 | $8,475,142 | $121,688 | $8,574,932 | $16,446,968 |
| 1998 | $1,048,777 | $7,864,873 | | $7,864,873 | $8,913,650 |
| 2000 | $6,766,074 | $12,245,739 | $59,808,260 | $72,037,938 | $78,804,013 |
| 2002 | $1,129,917 | $1,702,098 | | $1,702,098 | $2,832,016 |
| 2004 | $2,497,831 | $59,024 | $15,292,815 | $15,337,753 | $17,835,584 |
| 2006 | $8,311,596 | $11,809,417 | | $11,809,417 | $20,121,013 |
| 2008 | $5,652,425 | $14,354,805 | $35,944,872 | $50,284,861 | $55,937,287 |
| 2010 | $4,682,770 | $30,494 | | $30,494 | $4,713,264 |
| 2012 | $8,117,312 | $65,080 | $71,498 | $118,524 | $8,235,836 |
| 2014 | $1,375,765 | $81,408 | | $81,408 | $1,457,173 |
| 2016 | $3,586,714 | $3,699,887 | $27,709,099 | $27,724,723 | $31,311,438 |
| 2018 | $7,164,258 | $40,654,327 | | $40,654,327 | $47,818,586 |

Table 2: Cost of estimated ballots (in 2020 dollars) drawn to fulfill a ballot-level comparison audit with a risk limit of 10% for all house elections, senate elections, presidential elections in all states and D.C, elections for president and senate simultaneously, and for all federal races per year. Fixed costs of $10.85 million dollars nationwide are not show in this table.

otherwise, the top two candidates in the first round move to a second round special election. We only use data from the first round of votes. Similarly, Maine elected its congressional delegation using a ranked-choice voting (RCV) system. Under this new system one election [3] has advanced to an automatic runoff. Since RCV is an open research questions as it applies to RLAs, this election is omitted from the data. Second, in the two states in which presidential electors are awarded statewide and by congressional district, we consider only the statewide tally.

This data is limited in that it does not have results per congressional district for senate and presidential races. This complicates simultaneous audits for all federal races in a given state; as such, we do not model simultaneous audits between the house and other races.

## 3.2 Ballot-Level Comparison RLA

We model ballot-level comparison audits based on methods established by Stark in [12, 24]. Following these papers, we use $\alpha = 10\%$, $\gamma = 110\%$, and $\lambda = 50\%$. The sample-size multiplier $\rho$ is calculated as

$$\rho = \frac{-\log \alpha}{\frac{1}{2\gamma} + \lambda \log(1 - \frac{1}{2\gamma})}.$$

$\rho$ can be computed once and used as a constant value for all audits. A ballot-level comparison audit of one race is equivalently to a simultaneous ballot-level comparison audit of two of more races except in how the margin is computed. For a single audit, the margin $m$ is computed as the difference in votes between the top two candidates divided by the total number of votes cast. For a simultaneous audit the diluted margin $\mu$ is defined as the smallest difference, of all races on a ballot, between top two candidates divided by the total number of votes cast in any race. For simplicity, both types of margin will be referred to as $\mu$. Let $o_1$ and $u_1$ be the number of one-vote overstatements and understatements that occur in a sample of ballots of size $n$. Based on Stark's estimates the expected value of $o_1$ and $u_1$ is $.01 \times n$. Similarly, let $o_2$ and $u_2$ be the number of one-vote overstatements and understatements that occur in $n$ ballots. The expected value of $o_2$ and $u_2$ is $.001 \times n$.

To begin the audit an initial sample of $n = \lceil \frac{\rho}{\mu} \rceil$ ballots is drawn. The audit must continue if, for this sample, $\lambda\mu \leq o_1$ or $o_2 > 0$. In the case the audit continues after an initial sample is drawn, the size of the audit increments by one ballot until $\alpha$

[3] ME-02 (2018)

| Year | House | Senate | President | Total |
|---|---|---|---|---|
| 1976 | 3,056,933 | 36,021 | 6,017,929 | 9,110,882 |
| 1978 | 1,478,925 | 2,187,425 | | 3,666,349 |
| 1980 | 2,294,208 | 1,166,131 | 5,232,112 | 8,692,450 |
| 1982 | 2,812,383 | 91,409 | | 2,903,791 |
| 1984 | 2,763,789 | 1,334,943 | 2,113,955 | 6,212,686 |
| 1986 | 1,336,778 | 1,642,574 | | 2,979,351 |
| 1988 | 1,856,508 | 457,867 | 131,638 | 2,446,013 |
| 1990 | 1,692,934 | 32,987 | | 1,725,921 |
| 1992 | 3,686,002 | 1,598,891 | 448,850 | 5,733,742 |
| 1994 | 3,273,378 | 59,918 | | 3,333,295 |
| 1996 | 2,913,652 | 1,794,467 | 727,616 | 5,435,734 |
| 1998 | 716,862 | 1,624,494 | | 2,341,356 |
| 2000 | 1,784,375 | 2,545,204 | 12,158,854 | 16,488,432 |
| 2002 | 922,832 | 369,015 | | 1,291,846 |
| 2004 | 803,282 | 519,633 | 4,078,411 | 5,401,325 |
| 2006 | 2,968,985 | 2,805,489 | | 5,774,473 |
| 2008 | 2,527,934 | 3,222,366 | 15,072,271 | 20,822,570 |
| 2010 | 3,118,230 | 64,772 | | 3,183,002 |
| 2012 | 3,230,952 | 400,342 | 167,573 | 3,798,866 |
| 2014 | 1,822,241 | 209,509 | | 2,031,750 |
| 2016 | 2,026,415 | 820,667 | 5,938,308 | 8,785,388 |
| 2018 | 4,125,782 | 8,244,662 | | 12,370,444 |

Table 3: Estimated total ballots drawn to fulfill a ballot-polling audit with a risk limit of 10% for all house elections, senate elections, presidential elections in all states and D.C, and for all federal races per year.

is greater than the Kaplan-Markov MACRO P-value $P_{KM}$ of the sample. With the total error across $n$ ballots $U$ defined as

$$U = \frac{2\gamma}{\mu},$$

an upper bound on $P_{KM}$ of a sample of size $n$ is given as

$$P_{KM} \leq \frac{(1-\frac{1}{U})^n}{(1-\frac{1}{2\gamma})^{o_1}(1-\frac{1}{\gamma})^{o_2}}.$$

.

This is equivalent to drawing until

$$n \geq \frac{\rho + 1.4(o_1 + 5o_2 - 0.6o_2 - 4.4u_2)}{\mu}$$

We model house elections as a single audit of the one race for each congressional district. We model elections for the president and senate as both single audits and simultaneous audits of the two races.

### 3.3 Ballot-Polling RLA

To model ballot polling, we use an open-source risk-limiting audit tool Arlo [26]. Arlo implements BRAVO ballot-polling, a method created by Lindeman and Stark in 2012 [13], and was used by Michigan in a pilot audit of the 2020 Presidential Primary [20]. BRAVO takes the null hypothesis that the reported outcome is wrong, and then samples ballots until that null hypothesis is rejected at the chosen risk limit. In plurality contests with $n$ candidates, BRAVO tests $n-1$ null hypothesis, comparing each loser to the reported winner. In our modeling, we use a risk limit of $\alpha = 10\%$.

While simulating an audit of a contest, we proceed in rounds. Each round we sample without replacement a group of ballots distributed according to the reported results, then check the risk limit. Sample size is equal to the Average Sample Number (ASN),

$$\text{ASN} \approx 2\ln(1/\alpha)/x^2,$$

the expected number of ballots needed to reject the null hypothesis. The contest margin is $x$. We sample groups of ballots rather than individual ballots because logistically this is simpler for election officials, especially when sampling across jurisdictions for federal races. Each presidential, senate, or house race was audited separately, and the final ballot count from each audit averaged across 50 simulations. In close-margin contests where the sample size required to give a 90% chance of audit completion would be larger than $25,000$ ballots, Arlo falls back to a full recount.

| Year | House | Senate | President | All |
|---|---|---|---|---|
| 1976 | $15,040,110 | $177,223 | $29,608,210 | $44,825,539 |
| 1978 | $7,276,311 | $10,762,131 | | $18,038,437 |
| 1980 | $11,287,503 | $5,737,364 | $25,741,991 | $42,766,854 |
| 1982 | $13,836,924 | $449,732 | | $14,286,651 |
| 1984 | $13,597,841 | $6,567,919 | $10,400,658 | $30,566,415 |
| 1986 | $6,576,947 | $8,081,464 | | $14,658,406 |
| 1988 | $9,134,019 | $2,252,705 | $647,658 | $12,034,383 |
| 1990 | $8,329,235 | $162,296 | | $8,491,531 |
| 1992 | $18,135,129 | $7,866,543 | $2,208,342 | $28,210,010 |
| 1994 | $16,105,019 | $294,796 | | $16,399,811 |
| 1996 | $14,335,167 | $8,828,777 | $3,579,870 | $26,743,811 |
| 1998 | $3,526,961 | $7,992,510 | | $11,519,471 |
| 2000 | $8,779,125 | $12,522,403 | $59,821,561 | $81,123,085 |
| 2002 | $4,540,333 | $1,815,553 | | $6,355,882 |
| 2004 | $3,952,147 | $2,556,594 | $20,065,782 | $26,574,519 |
| 2006 | $14,607,406 | $13,803,005 | | $28,410,407 |
| 2008 | $12,437,435 | $15,854,040 | $74,155,573 | $102,447,044 |
| 2010 | $15,341,691 | $318,678 | | $15,660,369 |
| 2012 | $15,896,283 | $1,969,682 | $824,459 | $18,690,420 |
| 2014 | $8,965,425 | $1,030,784 | | $9,996,210 |
| 2016 | $9,969,961 | $4,037,681 | $29,216,475 | $43,224,108 |
| 2018 | $20,298,847 | $40,563,737 | | $60,862,584 |

Table 4: Cost of estimated ballots (in 2020 dollars) drawn to fulfill a ballot-polling audit with a risk limit of 10% for all house elections, senate elections, presidential elections in all states and D.C, and for all federal races per year.

### 3.4 Estimating Cost

Determining costs for conducting risk-limiting audits has proven to be extremely difficult given the nature of the procedure – the number of ballots reviewed is highly dependent on the margin of the contest. As a result, as noted in several pilot RLAs, the planning and budgeting cannot begin until the day after the election when the results from the election day and early voting are known. To account for this, we used simplifications in three different categories that encapsulate the audit burden – counting, cost, and workload [3].

First, regarding counting, or the number of ballots to be examined, we utilized the results of the ballot-polling and ballot-comparison models as described in previous sections to predict the expected sampled ballots in elections dating back to 1976. We then synthesized this information into two data points, (1) the average ballots to examine in President, Senate, and House races from 1976-2018 and (2) the ballots to examine in the most recent federal races (2016 President, 2018 Senate, and 2018 House). This split allowed us to diminish the effect of outliers in the recent federal races (2016 Wisconsin Presidential race, for example) while also understanding present-day voting trends. Additionally, because the results of RLA modeling were broken down by state, we were able to note specific states that overwhelmingly bore the burden of votes sampling more than others.

Second, regarding cost, or all direct and indirect expenses, we surveyed the budget sheets of several pilot RLAs conducted over the past decade to find necessary technology and materials. Several trends emerged, the majority of which were shared expenses among both ballot-polling and ballot-comparison audits - dice for random sampling, scales for counting ballots, ballot storage boxes, and document cameras. Given that most of these materials are relatively inexpensive and may already be owned by jurisdictions, we decided to disregard them in our estimation. We did, however, decide to factor in the cost of ballot scanners for ballot-comparison audits, as most states have varying availability of the required CVRs. Initially, we assumed every United States county would need to purchase ballot scanners to accommodate, but in seeing inflated fixed costs, we determined a more accurate estimation would allocate ballot scanners on a state-by-state basis. According to "The Administration of Cast Vote Records" by Kuriwaki, only 10 states collect CVRs from counties, for which we deducted our scanner allocation in our estimation [11]. For the other 40 states, we extrapolated results from the 2018 Michigan Pilot RLA, in which they noted 20% of ballots across sampled counties had a usable CVR. For the scanner itself, we found that most states use DS200 scanners for in-person ballot scanning, but several pilots have found that its inability to imprint sampling numbers to be insufficient for ballot-comparison audits [10]. DS850

| Race | Ballot-polling votes | Ballot-polling cost | Ballot-level comparison votes | Ballot-level comparison cost |
|---|---|---|---|---|
| House | 2,327,880 | $11,453,174 | 982,672 | $4,874,055 |
| Senate | 1,419,490 | $6,983,892 | 1,356,817 | $6,729,813 |
| President | 4,735,228 | $23,297,325 | 3,876,447 | $ 19,227,177 |
| President & Senate | | | 3,259,674 | $16,167,985 |
| All | 8,482,600 | $41,734,392 | 424,234 | $21,042,040 |

Table 5: Average of expected RLA sample size from 1976-2018 (data per year can be seen in Tables 1 and 3, and cost per year in Tables 2 and 4). We use a cost of $4.92 per ballot-polling vote and $4.96 per ballot-polling vote. Fixed costs of $10.85 million dollars for ballot-level comparison audits nationwide are not show in this table. In cases where the sum of ballots expected to be audited for senate and president was less than a modeled simultaneous audit for a given state, the better case was consider for the President-Senate results as well as to the national total.

central scanners have proven to be much more effective, but due to its bloated cost ($100,000), we used the cost of DS200 scanners ($5,300) for estimation with the assumption that new editions of the scanners would have imprinting capabilities (the producer of DS200 scanners has indicated as such) [7].

Third, regarding workload, or the person-hours invested in the audit, we again looked to reports from pilot RLAs to generate an accurate approximation. Indications of the time taken to examine a specified number of ballots, for both RLA methods, were extremely helpful to note how improvements in technology over the past decade have improved RLA efficiency. For example, the 2011 RLA pilots in California consistently reported a review time of around 90 seconds per ballot and a scan time of 20 minutes per ballot. By the 2016 Maryland and 2018 Michigan pilots, review times had shrunk to 40 seconds per ballot and an incredible scan time of 10 seconds (thanks in large part to DS850 scanners), values we would ultimately use for our cost estimation [10, 18]. Even so, we realized that reviewing and scanning did not incorporate the entire workload of an RLA audit; there were still outstanding labor expenses related to transportation, creating ballot manifests, handling boxes, and pulling samples. For this, we relied upon the 2019 Rhode Island pilot RLA once again, which detailed specific hours per ballot for these additional considerations. Given that Rhode Island is one of the smallest states by population and by size, we increased the hours per ballot value to extrapolate to the rest of the country, settling on 0.0001, 0.0004, 0.2, 0.035 hours/ballot for each of the extra labor expenses mentioned above, respectively. Altogether, this amounted to 0.246 labor hours/ballot for ballot-polling audits and 0.248 labor hours/ballot for ballot-comparison audits. Multiplying by a fixed labor cost per hour of $20 – accommodating for lower wages of temporary workers and higher wages of election officials – we found that the overall cost per ballot for ballot-polling audits to be $4.92 and for ballot-comparison to be $4.96 [21]. It must be noted that several assumptions were made in this calculation should be evaluated for use on state-by-state bases, but overall, we believe this to be a close estimate when averaging nationwide values.

# 4 Results

## 4.1 Expected Ballots Drawn

For either type of RLA, it is expected less than 4% of total ballots cast nationwide in a race must be sampled to meet a risk limit of 10% in nearly all cases. Shown in figure 1, ballot-level comparison audits estimate the house requiring less than 2% of the vote to be sampled, less than 4% of the senate vote to be sampled and less than 6% of the vote for president to be sampled. There are few notable outliers in the 2018 senate elections and 2000 presidential elections. Shown in figure 2, ballot-polling audits estimate the house requiring less than 4% of the vote to be sampled, less than 3% of the senate vote to be sampled and less than 6% of the vote for president to be sampled. Over the period, there is less variability for congressional seats than for the presidency in both types of audit.

We show results per year for both audits in tables 1 and 3. There are some cases in which separate audits of the senate and president are expected to draw less ballots together than run as one simultaneous audit. In this case we consider the best case.

The audit burden is not distributed evenly among the states. During the period we analyzed, most presidential elections were decided by a handful of "swing states.". While few states had competitive presidential election in a given year, many states have a competitive house elections. These types of races are both clearly displayed in figures 3 and 4 as outliers in the ballots drawn during an audit. In 2016, California, Michigan, and New Hampshire accounted for 93% of the ballot-comparison ballots drawn, and 76% of the ballot-polling ballots drawn, despite only making up 15% of the total ballots cast [14] across all federal elections.

We also show on average that less ballots are drawn if a ballot-level comparison audit were conducted than a ballot-polling audit. Table 6 shows the percent difference between the number of ballots drawn by each type of audit in each year of our data-set. Ballot-comparison almost always draws fewer ballots, sometimes dramatically so. For example, in 2014, conducting ballot-polling audits nationwide would have a

591% increase over ballot-comparison. On average, ballot-polling draws 114% more ballots. The implications of this are that over the period, ballot-comparison will draw significantly fewer ballots, saving time and money if implemented instead of ballot-polling.

## 4.2 Cost Estimation

After gathering the expected ballots examined for both types of audits from 1976-2018, we first looked to find the costs of the most recent federal elections. These elections, ranging from 2016-2018, included several races with incredibly close margins that inflated the examine vote count. As a result of these outliers, we have indicated the sum of these races to be an "upper bound" as shown in tables 2, 4 and 5. Another important distinction to highlight for the following charts is that we are estimating the cost of conducting an RLA in every state, not just ones with close margins who would be more likely request an audit. We envision this estimation to be used for the purposes of ultimately implementing a nationwide RLA policy, in which case we are establishing a ballpark range of potential allocation requirements.

As mentioned previously, we used $4.92 as the per ballot cost for ballot-polling audits and $4.96 as the per ballot cost for ballot-comparison audits. There were no additional fixed costs for ballot-polling audits, but using the 80% need for CVR-enabling ballot scanners in 40 states, the estimated fixed costs for ballot-comparison audits amounted to $10.85 million. A key assumption to clarify is the assignment of one ballot scanner for every county in the prescribed 40 states (2,047 counties), which follows the methods of the 2018 Michigan RLA pilot. Alternative allocation mechanisms could be utilized to cut costs in future estimations.

Several noteworthy observations stick out, the first being that conducting ballot-comparison RLAs on these sets of elections would have been cheaper than conducting ballot-polling RLAs, even with the fixed cost of allocating ballot scanners included (Tables 5 and 6). This result follows the reason why pilot RLAs reports have recommended ballot-comparison RLAs – they sample significantly fewer ballots, which in turn, lowers the cost for the RLA. Additionally, by looking closer at the expected ballots data set, we found that just a few outlier states with close margins ultimately drive the costs for a nationwide RLA; for example, Michigan accounts for 80% of the ballots sampled in the 2016 presidential race and Florida accounts for 98% of the ballots sampled in the 2018 senate races. This proves the point that budgeting for RLAs is extremely difficult due to the unpredictability in margins, but by setting a margin threshold for an immediate full hand recount (many states currently have this in place), a great portion of RLA expenses can be saved. Finally, part of our expected ballot modeling looked at conducting simultaneous audits – in this case presidential and senate – but due to different states with close margins in 2016, this method proved to be nearly two times more expensive than conducting the audits separately.

To understand if we were on the right track with this estimate, we looked into recent election audits with confirmed cost and came across a rather infamous example – Jill Stein's request for a full hand recount in Wisconsin after the 2016 presidential election, ultimately costing her campaign $2 million [25]. Looking at the results of our model, we estimate that 107,220 votes would be sampled for a ballot-polling RLA and only 3,803 votes for a ballot-comparison RLA. This amounts to $527,000 and $326,000 for ballot-polling and ballot-comparison (cost of ballot scanners included) RLAs, significantly less than the cost of a full hand recount, further supporting the purpose of an RLA. With our methods confirmed, we then aimed to account for the outliers by averaging the expected ballots to be sampled in all federal races from 1976-2018.

Once again, ballot-comparison RLAs prove to be more cost-efficient for the "average" election (table 5), one that assumes no state will conduct an RLA that leads to a full hand recount. Ballot-comparison audits especially reduce cost and workload with House races, as they are more likely to be non-competitive and stop after just tens of ballots are examined. One interesting observation in this result is that a ballot-comparison RLA would be more expensive if the fixed cost of ballot scanners were to be included, however we decided to amortize this cost over several election cycles to provide a better indicator of what an average election would cost.

# 5 Conclusion

Both ballot-level comparison and ballot polling RLA are less expensive and more effective than current methods. Based on these results, we recommend a federal RLA policy as follows. While the federal government would provide the funding for mandated risk-limiting audits on federal elections, the actual work of any audit is performed by each state. So a policy requiring risk-limiting audits would need to make payments to states. Complicating the distribution of funds is the variable audit burden each election, and the problem that every state wants to receive at least some federal funding. Audit policy should follow previous election law in its funding distribution schemes and ensure each state's audit costs are paid for.

In 2002, the United States Congress passed the Help America Vote Act (HAVA) with broad bipartisan support [1]. This act overhauled the nation's electoral process and created the Election Assistance Commission (EAC), which is responsible for making payments allocated under HAVA to states. One of HAVA's provisions required the replacement of certain types of voting machines. The funding states received for that purpose included both a minimum amount, which helped get the bill through Congress, and an amount proportional to their voting age population. Then in 2019, Senator

| Year | Ballot-Comparison | Ballot-Polling | Percent Difference |
|---|---|---|---|
| 1976 | 6,651,146 | 9,110,882 | 37.0 % |
| 1978 | 4,206,081 | 3,666,349 | -12.8 % |
| 1980 | 7,792,315 | 8,692,450 | 11.6 % |
| 1982 | 661,272 | 2,903,791 | 339.1 % |
| 1984 | 4,990,260 | 6,212,686 | 24.5 % |
| 1986 | 1,910,958 | 2,979,351 | 55.9 % |
| 1988 | 991,002 | 2,446,013 | 146.8 % |
| 1990 | 401,527 | 1,725,921 | 329.8 % |
| 1992 | 4,999,866 | 5,733,742 | 14.7 % |
| 1994 | 1,366,955 | 3,333,295 | 143.8 % |
| 1996 | 3,315,921 | 5,435,734 | 63.9 % |
| 1998 | 1,797,107 | 2,341,356 | 30.3 % |
| 2000 | 15,887,906 | 16,488,432 | 3.8 % |
| 2002 | 570,971 | 1,291,846 | 126.3 % |
| 2004 | 3,595,884 | 5,401,325 | 50.2 % |
| 2006 | 4,056,656 | 5,774,473 | 42.3 % |
| 2008 | 11,277,679 | 20,822,570 | 84.6 % |
| 2010 | 950,255 | 3,183,002 | 235.0 % |
| 2012 | 1,660,451 | 3,798,866 | 128.8 % |
| 2014 | 293,785 | 2,031,750 | 591.6 % |
| 2016 | 6,312,790 | 8,785,388 | 39.2 % |
| 2018 | 9,640,844 | 12,370,444 | 28.3 % |

Table 6: Comparison of total ballots drawn to fulfill ballot-comparison and ballot-polling audits for all federal races per year, and the percent change between ballot-comparison and ballot-polling.

Wyden introduced the Protecting American Votes and Elections (PAVE) Act [2]. The PAVE act amends HAVA with requirements for paper ballots, election audits, and election cybersecurity. It empowers the EAC to reimburse states for costs of post-election audits and establish procedures for submitting eligible costs. However, it provides no guarantees that a state receives election-audit funding under this program.

We propose that funding is distributed by the EAC as described in the PAVE act, while also guaranteeing a minimum payment. Requiring states to submit costs for reimbursement ensures that each state's variable audit burden is covered. The minimum payment made to each state should be in proportion to its voting age population. We would like to recommend minimum payments be distributed on a basis of margin; but, swing-state status (or number of toss-up elections in a given state) is not constant over the period.

For recommendations on efficient state implementation, any policy should suggest a few optimal practices. First, emphasis should be placed on conducting ballot-level comparison audits, as it is more cost-efficient, makes it easier to trace discrepancies on a ballot-level basis, and ensures a simple transition from the full hand recount process that many states already have in place. Additionally, we advise that states conduct a pilot RLA to prepare election officials for the expected logistical planning of future election RLAs. Third, state election officials should look to conduct centralized audits (in one or two locations), so that the RLA process can be standardized. The costs of transporting the ballots may increase significantly in large states such as California and Alaska, but the guarantee of a legitimized RLA may outweigh the potential downside of expenses. Finally, we suggest that states look into conducting simultaneous risk-limiting audits when the margin for federal races is not particularly close, as the auditing burden can be notably reduced.

Further research on this topic should consider auditing the close races which decide the presidency and party control of either chamber of Congress. Since variable costs of auditing every federal race is concentrated in these states, we do not expect an overwhelming difference expected number of ballots drawn between this proposal and results presented here. However, auditing a fewer number of races will reduce previously mentioned non-variable RLA costs. Moreover, current RLA methods are only applicable for first-past-the-post single-member district elections. While multi-member districts seems unlikely for federal elections in the near future, it may be the case that other states join Maine in implementing a RCV method for congressional elections. Then, it would be necessary to develop additional RLA methodologies to cover these scenarios.

## Availability

The data we used and results we produced are available at https://github.com/jbrasch/rla. The scripts we used to produce results, figures in the paper, and additional figures are hosted there as well.

## References


[1] Help america vote act of 2002, h.r.3295, 107th cong. https://www.congress.gov/bill/107th-congress/house-bill/3295.

[2] Protecting american votes and elections act of 2019, s.1472, 116th cong. https://www.congress.gov/bill/116th-congress/senate-bill/1472.

[3] Jennie Bretschneider, Susannah Goodman, Mark Halvorson, Mark Lindeman Pam Smith, and Philip B. Stark. Risk-limiting post-election audits: Why and how (version 1.1), 2012. https://www.stat.berkeley.edu/~stark/Preprints/RLAwhitepaper12.pdf.

[4] MIT Election Data and Science Lab. U.S. House 1976–2018, 2017. https://doi.org/10.7910/DVN/IG0UN2.

[5] MIT Election Data and Science Lab. U.S. President 1976–2016, 2017. https://doi.org/10.7910/DVN/42MVDX.

[6] MIT Election Data and Science Lab. U.S. Senate 1976–2018, 2017. https://doi.org/10.7910/DVN/PEJ5QU.

[7] LLC Election Systems & Software. Cost Table, 2014. https://www.michigan.gov/documents/sos/ESSCost_549582_7.pdfU.

[8] Zachary Halaschak. Michigan to conduct 'zero-margin risk-limiting' audit in antrim county amid fracas over dominion machines. *The Washington Examiner*, 2020.

[9] Joseph Lorenzo Hall, Luke Miratrix, Philip B Stark, Melvin Briones, Elaine Ginnold, Freddie Oakley, Martin Peaden, Gail Pellerin, Tom Stanionis, and Tricia Webber. Implementing risk-limiting post-election audits in california. *State of California*, 2009. https://www.usenix.org/legacy/event/evtwote09/tech/full_papers/hall.pdf.

[10] Liz Howard, Ronald L. Rivest, and Philip B. Stark. A review of robust post-election audits: Various methods of risk-limiting audits and bayesian audits. *Brennan Center for Justice*, 2019. https://www.brennancenter.org/sites/default/files/2019-11/2019_011_RLA_Analysis_FINAL_0.pdf.

[11] Shiro Kuriwaki. The administration of cast vote records in U.S. states. 2020. https://www.shirokuriwaki.com/papers/kuriwaki_cvr-supp.pdf.

[12] M. Lindeman and Philip B. Stark. A gentle introduction to risk-limiting audits. *IEEE Security Privacy*, 10(5):42–49, 2012. https://www.stat.berkeley.edu/~stark/Preprints/gentle12.pdf.

[13] Mark Lindeman, Philip B Stark, and Vincent S Yates. Bravo: Ballot-polling risk-limiting audits to verify outcomes. *Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE)*, 2012. https://www.usenix.org/system/files/conference/evtwote12/evtwote12-final27.pdf.

[14] Michael P. McDonald. 2016 November General Election Turnout Rates, 2018. http://www.electproject.org/2016g.

[15] Emil Moffat. With biden ahead, georgia begins hand recount of nearly 5 million ballots. *National Public Radio*, 2020.

[16] Jennifer Morrell. Risk-limiting audits: Postelection audits, a summary. *National Conference of State Legislatures*, 2020. https://www.ncsl.org/research/elections-and-campaigns/risk-limiting-audits.aspx.

[17] Lawrence Norden and Wilfred U. Codrington III. America's voting machines at risk – an update. *Brennan Center for Justice*, 2018. https://www.brennancenter.org/our-work/research-reports/americas-voting-machines-risk-update.

[18] Maryland State Board of Elections. Post-election tabulation audit pilot program report. *State of Maryland*, 2016. https://www.elections.maryland.gov/press_room/documents/Post%20Election%20Tabulation%20Audit%20Pilot%20Program%20Report.pdf.

[19] California Secretary of State. Post-election risk-limiting audit pilot program 2011-2013: Final report to the united states election assistance commission. *State of California*, 2014. https://votingsystems.cdn.sos.ca.gov/oversight/risk-pilot/final-report-073014.pdf.

[20] Michigan Department of State Bureau of Elections. Risk-limiting audit pilot of the 2020 michigan presidential primary, 2020. https://www.michigan.gov/documents/sos/Michigan_RLA_Report_693501_7.pdf.

[21] Report of the Rhode Island RLA Working Group. Pilot implementation study of risk-limiting audit methods in the state of rhode island, 2019. https://verifiedvoting.org/wp-content/uploads/2020/07/RI-RLA-Report-2020.pdf.

[22] Philip B. Stark. Risk-limiting post-election audits, 2008. https://www.stat.berkeley.edu/~stark/Seminars/ksu08.pdf.

[23] Philip B Stark. Risk-limiting postelection audits: Conservative *p*-values from common probability inequalities. *IEEE Transactions on Information Forensics and Security*, 4(4):1005–1014, 2009. https://www.stat.berkeley.edu/~stark/Preprints/pvalues09.pdf.

[24] Philip B Stark. Super-simple simultaneous single-ballot risk-limiting audits. *Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE)*, 2010. https://www.stat.berkeley.edu/~stark/Preprints/superSimple10.pdf.

[25] Riley Vetterkind. Elections commission estimates potential recount would cost 4 times more than 2016's. *Wisconsin State Journal*, 2020.

[26] VotingWorks. Open-source risk-limiting audit software, 2020. https://github.com/votingworks/arlo.