City of Fairfax, VA

Pilot Risk-Limiting Audit

December 7, 2018

verifiedvoting

By Mark Lindeman, for the Verified Voting Foundation.

This report was made possible with financial support from Microsoft.

Table of Contents

In August 2018, with technical assistance from Verified Voting, the City of Fairfax, Virginia conducted the state's first pilot of risk-limiting audits (RLAs). This pilot changed the policy discussion about RLAs in Virginia. It has lessons for election officials and policymakers throughout the country – and perhaps for RLA advocates as well.

Risk-limiting audits implement a simple, compelling idea: instead of relying on the accuracy of voting machines, check enough ballots by hand to obtain strong evidence that the declared winner(s) of each audited contest really got more votes. A recent consensus report of the National Academies of Science, Engineering, and Medicine declares that within a decade, "[r]isk-limiting audits should be conducted for all federal and state election contests, and for local contests where feasible."[1] But risk-limiting audits are widely perceived as complicated. Largely this is a function of unfamiliarity, compounded by the intrinsic complexity and variety of election practices that intersect with RLAs. Pilot audits cut through the mystery and noise, placing audit processes in the hands of election officials.

Evidence-based elections and risk-limiting audits: a brief introduction

The 2016 U.S. presidential election campaign foregrounded long-standing concerns about the accuracy of electronic vote tabulations. Computer scientists have warned that computerized voting and counting systems are vulnerable to error or malicious subversion, and must be checked using methods that do not rely on the correctness of hardware or software.[2] The U.S. intelligence community and other credible observers have reported on widespread cyberattacks on election systems during the campaign, including the data breach of a state voter registration database (widely reported, but not officially confirmed, to have been Illinois').[3] Officials emphasized that there was no evidence that any data had been changed, nor was there evidence that *votes* had been changed.

Unfortunately, due to poorly designed equipment and procedures, evidence that votes *hadn't* been changed was fragmentary. Tens of millions of Americans voted on systems that provide no verifiable record of their votes. Many more marked and cast their votes on paper ballots, but in states that do not systematically compare those ballots to the official returns. Paper ballots and systematic comparisons of paper to official returns are prerequisites of evidence-based elections.[4] A voting system that produces accurate results, but provides no way to know whether it did, is inadequate. It provides far too many ways for resourceful adversaries to undermine public confidence in election integrity.

[1] National Academies of Sciences, Engineering, and Medicine (2018). *Securing the Vote: Protecting American Democracy*, Washington, DC: The National Academies Press. https://doi.org/10.17226/25120.

[2] *Securing the Vote,* referenced above, provides an authoritative overview of these concerns.

[3] For instance, Adam Thorp, "Illinois election officials: 'Very likely' state was target of Russian hackers." *Chicago Sun-Times*, July 13, 2018, https://chicago.suntimes.com/news/illinois-election-officials-very-likely-state-was-target-of-russian-hackers/.

[4] Philip B. Stark and David A. Wagner, "Evidence-Based Elections," *IEEE Security and Privacy* 10 (2012), available at https://www.stat.berkeley.edu/~stark/Preprints/evidenceVote12.pdf.

The basic strategy for evidence-based elections can be summarized as follows: use paper ballots, protect them, and check them. More specifically:

1. Voters must vote on voter-marked paper ballots – either marked by hand or using ballot marking devices, but in either case, with a convenient and accessible means for voters to verify their ballots before officially casting them. Direct Recording Electronic voting machines that produce “voter-verifiable paper audit trails” provide, at best, an obsolescent stopgap: most voters never check them, and often they are hard to audit.

2. Voted paper ballots must be carefully stored and managed to ensure that no ballots are added, removed, or altered – and procedures should provide strong evidence that in fact the ballots were properly managed. The means used to protect ballots from tampering should be reviewed by security experts, and compliance audits should be performed to confirm, for instance, that ballot containers were properly secured.

3. Voted ballots also must be checked in robust post-election vote tabulation audits, in which audit judges manually review a random sample of voted ballots (and possibly additional ballots) and compare them to the reported results. As far as is feasible, these audits should be risk-limiting audits. Informally speaking, if an election outcome (e.g. the announced winner) is wrong, a risk-limiting audit is very likely to correct it through a full hand count. Recounts, as well as risk-limiting audits, should rely on human inspection of the actual voted ballots.

More formally, a risk-limiting audit (RLA) provides a large, prespecified minimum chance, if a reported outcome[5] for an audited contest is incorrect – i.e., disagrees with what an accurate full hand count of the ballots would show – of leading to a full hand count that corrects the outcome. (Legally, the full hand count might be part of the audit, or it might be a separate recount required based on the audit findings.) The risk limit is the corresponding small maximum chance that an incorrect contest outcome will *not* be corrected. For instance, if the audit has at least a 95% chance of correcting an incorrect outcome, it has a 5% risk limit. The actual chance of correcting a wrong outcome may be much larger – depending, for instance, on the actual margin of victory. If a risk-limiting audit with a small risk limit does *not* lead to a full hand count, that constitutes strong evidence that the reported outcome is correct based on the ballots.

Risk-limiting audits have often been mythologized as “statistical methods” that rely on mathematical expertise. The implication is that most people cannot understand or perform them. This view is mistaken, akin to describing public buildings as “engineering models” because engineers participate in their design. A risk-limiting audit is a routine post-election process, generally performed by election officials and open to public observation.

[5] The outcome is the legal and official consequence of an election: for instance, who will take office, who will participate in a runoff election, or whether a referendum will pass.

Risk-limiting audits are adaptable: they can be conducted in a variety of ways, some of which are described later in this report. And they can be highly efficient, in the sense that they can be designed to do only as much work as necessary to confirm an election outcome – but no less. Note that risk-limiting audits are designed to attain risk limits for particular, specified contests. Risk-limiting audits of some contests can be combined with less stringent, but still valuable, audits of other contests.

In November 2017, after a series of pilot audits dating back to 2010, Colorado became the first state to conduct risk-limiting audits statewide. In November 2018, Colorado conducted the first risk-limiting audits of statewide contests. Rhode Island has a statutory requirement to conduct risk-limiting audits beginning with the 2020 presidential preference primary. State laws in California and Washington explicitly welcome pilot or voluntary risk-limiting audits, and local election officials in other states have conducted pilot audits on their own authority. And in 2017 the Virginia General Assembly amended its audit law, adding a requirement to conduct "risk-limiting" post-election audits annually, effective July 1, 2018.

Risk-limiting audits in Virginia: the background

Like many states, Virginia has a part-time state legislature. The state constitution specifies that the regular session of the General Assembly is no longer than 60 days in even-numbered years, and 30 days in odd-numbered years, unless supermajorities of both houses vote to extend it.[6] (For their service, members are paid about $18,000 per year plus a modest per diem.[7]) This abbreviated calendar means that legislation must move quickly or not at all – and it does not facilitate careful policy assessment.

Arguably Virginia's current post-election audit law offers a case in point. In 2017, the Virginia General Assembly amended its audit law, adding a requirement to conduct "risk-limiting" post-election audits annually, effective July 1, 2018. The main provisions are as follows (additions to the previous law in italics, deletions in strikethrough):[8]

A. The Department of Elections shall be authorized to coordinate a post-election risk-limiting audit annually of ballot scanners in use in the Commonwealth. The localities selected for the audit shall be chosen at random with every locality participating in the Department's annual audit at least once during a five-year period. The purpose of the audits shall be to study the accuracy of ballot scanner machines.

B. No audit conducted pursuant to this section shall commence until after the election has been certified and the period to initiate a recount has

[6] Constitution of Virginia, Article IV, Section 6, https://law.lis.virginia.gov/constitution/article4/section6/.

[7] National Conference of State Legislatures, "2017 Legislator Compensation Information," http://www.ncsl.org/research/about-state-legislatures/2017-legislator-compensation-information.aspx.

[8] This text can be found at https://lis.virginia.gov/cgi-bin/legp604.exe?171+ful+CHAP0367. The legislative history and links to all versions of the bill and to fiscal impact statements can be found at https://lis.virginia.gov/cgi-bin/legp604.exe?171+sum+SB1254.

> expired without the initiation of a recount. An audit shall have no effect on the election results.

As other observers have noted, much of this language is problematic. The word "risk-limiting" was inserted by Virginia's Senate after the bill was introduced, and the concept is not well integrated into the statute.[9] A risk-limiting audit, by definition, must be able to proceed to a full hand count if necessary to correct contest results. To begin the audit after the recount deadline defeats this central purpose. Moreover, the focus on "study[ing] the accuracy of ballot scanner machines," echoed in the reporting requirements,[10] is misdirected: risk-limiting audits check election outcomes and overall tabulation accuracy, not necessarily the accuracy of individual scanners.

The shift in state officials' role, from "conduct[ing]" to "coordinat[ing]" the audit, also was consequential. In Virginia, elections are conducted by local election officials in 95 counties and 38 independent cities. Each of these 133 localities has a chief election official – usually a general registrar – and an electoral board tasked with supervising the general registrar and election staff. In the short run, the shift may have been perceived as an expedient way to avoid state budget impacts.[11] But the requirement engendered skepticism – if not outright hostility – toward what some local election officials perceived as a pointless and burdensome unfunded mandate.

Other requirements also created difficulties. Virginia readily could have implemented constructive post-election audits[12] beginning in mid-2018, but to implement comprehensive risk-limiting audits in that time frame would be very ambitious -- particularly without additional funding. The requirements that localities be selected at random and that every locality participate during a five-year period tend to complicate implementation. Although risk-limiting audits should be based on random samples, there is no inherent need for those samples to include every locality on an arbitrary schedule. The language also suggests that voluntary audits may not satisfy the five-year participation requirement, thus tending to discourage local pilots.

[9] The legislative history, at https://lis.virginia.gov/cgi-bin/legp604.exe?171+amd+SB1254AS, shows that the word "risk-limiting" simply was added to the sentence.

[10] Subsection D requires the Department of Elections to report "a comparison of the audited election results and the initial tally for each machine audited." As discussed below, checking individual machine tallies is generally the least efficient audit method!

[11] The fiscal impact statement prepared for the Senate amended bill of January 17, in which ELECT would have conducted the audit, called for two full-time employees – a "voting equipment audit coordinator" and a "statistical analyst" – to implement the audit: https://lis.virginia.gov/cgi-bin/legp604.exe?171+oth+SB1254FE122+PDF. The House substitute bill of February 10 changed "conduct" to "coordinate," implying that local jurisdictions would conduct the audit. The subsequent impact statement states, "The proposed legislation is not expected to have a state fiscal impact as the bill no longer requires the Department of Elections to conduct a post-election risk-limiting audit of ballot scanner machines annually…." https://lis.virginia.gov/cgi-bin/legp604.exe?171+oth+SB1254FH1122+PDF It is not obvious why the state would need two full-time staff members to conduct, perhaps, one statewide RLA each year, but could coordinate plausibly dozens of local RLAs with no additional staffing.

[12] We will use "post-election audits" as shorthand for post-election *vote tabulation* audits, whether or not they are risk-limiting. Other kinds of audits can (and should) also be conducted after elections.

These considerations illustrate some of the difficulties in crafting reasonable statutory language that will lead to the desired result: verified election outcomes. Fuller consultation with state and local election officials, as well as specialists in post-election audit legislation, might have produced legislation that would set a clearer direction and be easier to implement. That said, Virginia's statute has created an impetus to implement risk-limiting audits, while allowing considerable discretion in the early phases of implementation. Election officials have expressed interest in seeking revisions in the future.

Rapid development of the Fairfax pilot

In this context, Verified Voting met with Virginia's newly-appointed Commissioner of Elections, Chris Piper, and Confidential Counsel James Heo in May 2018 to discuss the possibility of helping Virginia begin to implement risk-limiting audits in Virginia under the current law. Throughout the country, Verified Voting works to assist election officials and other stakeholders in implementing RLAs and other election procedures that enhance election security. We believe that well-designed RLAs benefit election officials, as well as other citizens, because they address public concerns about election integrity and allow officials to learn more about the performance of several key election processes. In addition, well-designed RLAs can persuade losing candidates that the outcome is accurate, in some cases preventing them from insisting on a full recount. More efficient audits benefit everyone: they reduce burdens on election officials and audit staff, allow them to work more carefully, and are easier for interested citizens to observe, enhancing their benefits for public confidence.

In Verified Voting's view, and as the experience of Colorado underscores, pilot audits can be an invaluable step in implementing full-scale RLAs. Pilot audits offer a variety of benefits. Pilots enable election officials to experiment with new procedures on a small scale. Often these pilots can be conducted outside the ordinary election calendar, removing artificial time constraints. (In this respect, the statutory requirement to conduct RLAs after the recount deadline – paradoxical as it is – did facilitate pilots.) Pilots demystify RLAs, allowing election officials to approach them not as an arcane innovation, but as a series of specific tasks that they can plan for, and often find better ways to do. Moreover, pilots can create the common ground upon which election officials and specialists collaborate in designing the best possible audits. Indeed, pilots can enable election officials to move from cautious skepticism about RLAs to enthusiastic support, as they begin to see the benefits.

In this and subsequent conversations, we found that officials at the Virginia Department of Elections (often called ELECT) had a similar orientation. They faced a statutory mandate, and they were more than willing to work with local election officials and helpful outside experts to ease the implementation.

A pilot could only succeed with leadership from a fully committed local election official – one not just willing to endure unfamiliar tasks, but eager to dive into implementation details. Fortunately, Brenda Cabrera, General Registrar and Director of Elections in the City of Fairfax, Virginia, and the city's Electoral Board were up to the effort. (The City of Fairfax is not to be confused with Fairfax County, which surrounds it.) Liz Howard, who had served at ELECT through 2017 and now works for the Brennan Center, introduced the Verified Voting team to Cabrera. On May 22, Cabrera and Curt Chandler, the chair of Fairfax's electoral board, participated in a conference call with Verified Voting representatives and others about a possible pilot audit of the July primary election. Despite some notes of caution, Cabrera and Chandler expressed interest in further discussions. Remarkably, just over two months later, the city completed its pilot, in full collaboration with state ELECT officials, and with dozens of local election officials from around the state in attendance. Verified Voting provided extensive assistance in planning the basic audit sample design and provided software to support the audit.

Rapid development of this pilot was extraordinary. The basic proposal, and the city and state decisions to proceed with a pilot, were completed the week of June 18, less than a month after the initial discussion with City of Fairfax officials. The pilot was conducted six weeks later. This compressed schedule required an intense collaborative process. Many implementation decisions were made in a series of conference calls among Brenda Cabrera, James Heo, Verified Voting's Senior Science and Technology Officer, Mark Lindeman, and Advisory Board member and Audit Specialist, John McCarthy, with other state and local election officials when appropriate. Between the calls, staff members from Fairfax, ELECT, and Verified Voting all rapidly generated draft documents and software interfaces, answered each other's questions, and commented on each other's work.

Audit sample design options for Virginia

The pilot sought to inform broader discussions about how to implement risk-limiting audits statewide, in the near term and in the longer term. Those choices hinge on audit sample design: the basic plan(s) for sampling ballots and analyzing the results. Sample design is just one element of audit planning, but it determines the scope – both the work conducted and the benefits obtained – so it requires careful consideration.

There are three common methods for conducting risk-limiting (and other) post-election tabulation audits, which can be combined in a single pilot. For educational and policy reasons, state and local election officials decided that the pilot would test elements of all three. (Appendix 2 provides more information on each of these methods.) In brief:

- Ballot-level comparison examines a random sample of *individual ballots* and compares the audit interpretation of each ballot to its corresponding machine interpretation. Existing precinct-count voting systems generally do not support this kind of audit, so it was necessary to rescan and reinterpret the ballots.

- Ballot polling also examines a random sample of ballots, but instead of comparing to machine interpretations of individual ballots, it simply looks for a preponderance of votes for the reported winner (or outcome). Thus ballot polling is like an "exit poll" of actual ballots. This method is less efficient than ballot-level comparison, especially for small margins.

- Batch-level comparison examines the ballots in a random sample of batches for which machine counts are available (for instance, a batch may comprise the ballots cast in one precinct), and compares the audit counts to the machine counts. This method generally requires the most counting.

All these methods use some form of ballot manifest, essentially a table that details all the various ballots[13] cast in an election and where they are stored. For instance, the manifest might include columns for batch ID, number of ballots, and location (such as a particular storage box). The ballot manifest is crucial both for drawing a valid sample and for retrieving the ballots to be audited. Ballot manifests must not rely on the accuracy of the voting system: all counts should be checked by other means.

None of the basic methods – ballot-level comparison, ballot polling, or batch-level comparison – offers an optimal "turnkey" solution for risk-limiting audits in Virginia, but each can add value. In the long run, Virginia probably is well served to move toward ballot-level comparison. For the foreseeable future, it could reasonably decide to combine all three methods in various contexts.

Pilot sample design

Verified Voting therefore suggested, and election officials agreed, that the City of Fairfax pilot include some element of all three methods. This approach offered several potential advantages. It would help in breaking RLAs down into intelligible (even if in some cases unfamiliar) election processes and presenting choices among those processes. It would begin to expose and explore some of the logistics involved in implementing those processes in Virginia, thus informing future decisions and plans. In particular, it would provide an opportunity to time some processes. To be sure, piloting even one of these methods could achieve many of these goals.

In designing this multi-method pilot, the planners paid careful attention to ensuring that the workload would remain manageable. The context for these choices was a small and simple election: in the City of Fairfax, the June 12 primary covered just one party (Republican) and one contest (U.S. Senate), with under 950 ballots cast. Corey Stewart, who won the nomination statewide by about 1.7 percentage points, defeated Nick Freitas in Fairfax by a larger proportional margin of about 11 points: 439 votes to 337. A third candidate, E. W. Jackson, received only 159 votes, and 12 ballots were recorded as undervotes or overvotes.

This small election made a *ballot-level comparison* RLA limited to Fairfax unusually feasible. Specifically, it was feasible to rescan and retabulate all the ballots, using a commercial off-the-shelf (COTS) Visioneer Patriot H60 scanner provided by Verified

[13] In some jurisdictions where ballots routinely comprise more than one ballot card, it is appropriate to enumerate the cards. That complication is not applicable here.

Voting, rated at 65 pages per minute.[14] Using the retabulation results, the planners anticipated that the audit could confirm the "outcome" of the primary contest in Fairfax at a 5% risk limit, with some tolerance for error,[15] by auditing approximately 71 ballots. Although the "winner" of one city in a statewide primary is of no real consequence, treating this notional outcome as the RLA target provided an intuitive goal for this part of the audit, as well as a context for discussing how risk calculations could affect full statewide RLAs.

The *ballot-polling* audit posed a minor quandary. Because ballot-polling RLAs have unpredictable workloads, it would be hard to select a reasonable initial sample size that would provide a large chance of attaining a 5% or 10% risk limit in just one round of auditing. The team decided upon a fixed sample size of 300. This sample size was estimated to provide roughly a 50-50 chance of attaining a 10% risk limit, but its main rationale was to increase the work of ballot retrieval and audit adjudication substantially yet manageably. It was also considered that the contrast between ballot polling and ballot-level comparison could illustrate the benefits of moving toward systems that support ballot-level comparison without rescanning.

The *batch-level comparison* audit was the simplest aspect of the pilot. A batch-level RLA at the precinct level would have entailed hand-counting all or almost all the precincts, which was out of the question.[16] The team decided to hand-count just one precinct, which could plausibly represent the city's share of a statewide RLA in a close but not razor-close election. Considering the statewide margin of 1.7 percentage points, an actual statewide RLA at a 5% risk limit might have entailed auditing less than 200 batches statewide, out of over 2400 precincts, so auditing no more than one of the city's seven precincts was reasonable.

For simplicity, it was predetermined to hand-count the Central Absentee Precinct (CAP) on the first day of the audit, while ballots were being divided into batches and rescanned, instead of randomly selecting a precinct at the end of that day. Choosing the CAP minimized the counting: the CAP contained just 61 ballots, while the election day precincts had counts ranging from 110 to 187. This choice was a concession to concerns about delaying the other processes. At the same time, it seemed most likely to expose unusually marked ballots to scrutiny, because absentee voters in some ways are at greater risk of marking their ballots incorrectly. In actual RLAs, jurisdictions cannot choose

[14] The scanner can scan either simplex or duplex at the same speed, although postprocessing time can vary. We elected to ensure that the ballots remained "right side up" and to scan just one side. No marks were observed on the backs of the ballots, apart from the preprinted electoral board seal.

[15] Specifically, this sample size was chosen to allow for one one-vote overstatement, such as a ballot initially counted as an undervote that the adjudicators interpreted as a vote for the runner-up, thus reducing the margin by one vote.

[16] Because the retabulation batches were smaller than the original precincts, a batch-level comparison audit based on the retabulation results would have been somewhat less burdensome – but no value was seen in testing this approach. Attaining risk limits was never a central goal of the pilot, and this batch-level method seems unlikely to be used in practice: If an elections office is willing to rescan all the ballots in order to conduct an audit, it probably can conduct any additional work needed to support ballot-level comparison, which is far more efficient. This analysis exemplifies how the pilot's broader goals informed its design.

which precincts or ballots to audit – although, ideally, they can choose to do additional auditing on their own initiative.

Physical logistics

Most of the logistical specifics described in this and the following section were specified in writing, both to control the pilot process and to inform future implementation efforts.[17]

Voted ballots in the City of Fairfax are maintained by state circuit court officials at the Fairfax County Courthouse. To conduct an audit, the state commissioner of elections, Chris Piper, had to make a formal request to unseal the ballots for that purpose. The county courthouse is just two blocks from the elections office, and provided a far better venue for a public event. Accordingly, Brenda Cabrera arranged with court officials to conduct the audit in a large jury room at the courthouse, which could accommodate several working areas and many observers.

Chain of custody was carefully maintained. Each morning of the audit, court officials moved the ballots to a locked room adjoining the jury room, where the ballots initially remained in the custody of the Clerk of the Circuit Court. From there, sworn election officers who participated in the audit signed out ballots as needed, returning them to court custody once no longer in use. Only these officers and other election officials, including the Electoral Board members, were permitted to handle ballots while they were checked out. A chain-of-custody form was designed and used for the express purpose of signing ballots out and in.

The audit was scheduled to occur in the jury room over three days, Wednesday through Friday (August 1-3, 2018). On Wednesday, people involved in the audit refined the physical layout and tested equipment and software using facsimile ballots. Thursday – the first public day, and the first involving voted ballots – was spent rescanning and retabulating ballots in small batches, conducting the batch comparison hand-count of the Central Absentee Precinct, and using a public ceremony to generate a random seed from which the ballot samples were generated. On Friday, the ballot samples were retrieved and adjudicated, and results were announced.

Retrieving specific ballots was a challenge because the voted ballots, quite properly, were not labeled with identifiers. Ballot-polling audits are not systematically affected by innocent errors in ballot retrieval, but ballot-level comparison audits depend on accurately matching individual ballots to their corresponding individual Cast Vote Records. We considered physically adding ID numbers to voted ballots, either using a printer in the courthouse or manually with a colored pen. (The IDs could have been added on the backs of the ballots.) However, these approaches might trouble observers, could be difficult to apply on a larger scale, and were not unambiguously compliant with Virginia election law.

[17] ELECT's report on the pilot, available at https://www.elections.virginia.gov/Files/Media/Agendas/2018/20180920-RLA_Report.pdf, documents the procedures in further detail.

Instead, on Thursday, the ballots were manually divided into batches of 25, except for the last batch in each precinct. This work, as well as subsequent ballot retrieval on Friday, was performed at three stations by teams of two sworn election officers each. These six officers had worked as pollworkers on election day, and were paid modest per diems.[18] The batching procedure yielded 41 batches in all. Each ballot was identified by – but not imprinted with – a ballot ID consisting of the precinct identifier, the batch number within that precinct, and the sequential ballot number within that batch, as in **P1-4-019** for the 19th ballot in the fourth batch in precinct 1.[19] Each batch of ballots was kept in a labeled folder, which election officers annotated with the number of ballots contained in it. Thus, on Friday, a sampled ballot with a given ID could be retrieved by finding the correct folder and then counting down to the appropriate ballot. After retrieval, the unused ballots were returned to court custody; the retrieved sample ballots were sequestered in a separate container.

Both scanning and adjudication were conducted at a work table across from the ballot batching and retrieval stations. The audit software ran on a Windows laptop. During scanning, three officials worked together to keep track of which batch was being scanned, to ensure that the software correctly identified the batch, and to manage ballots in the scanner. (The Visioneer scanner was connected to the laptop via USB port.) A digital projector allowed observers to follow the progress of the scanning. During adjudication, the software displayed the ID of each ballot to be adjudicated – to be compared with the ID on the cover sheet identifying the ballot –and allowed adjudicators to record their interpretation of the vote(s) recorded on the ballot (see photo on page 15). An analog document projector allowed observers to see the ballots, while the digital projector let them watch as an election official entered the adjudications on the computer. After adjudication, the container with retrieved ballots was sealed and returned to court custody.

Additional audit logistics

The batch-level comparison audit of the Central Absentee Precinct, as mentioned above, happened just before the absentee ballots were divided into batches of 25. The two sworn election officers were instructed to take turns hand-tallying the 61 ballots in the CAP, each closely observed by the other, and then to compare results.

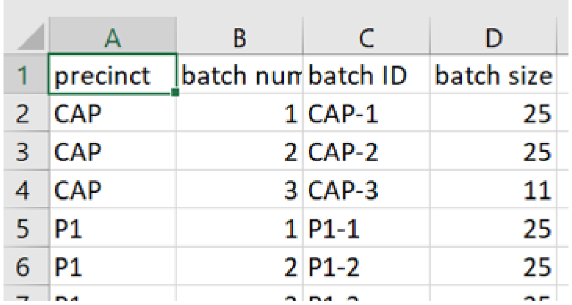
After each batch of ballots was scanned, any ballots identified by the software as undervotes or overvotes were manually reviewed by officials, using the same software adjudication interface to enter their interpretations as was used when reviewing ballots in the audit sample. Any ballots that were unscannable, presumably due to damage, would

[18] The team assigned to divide the Central Absentee Precinct into batches also performed the hand count. To balance workload across teams, this team was assigned the two smallest election day precincts; the other two teams worked with two larger precincts each.

[19] An audit observer suggested that it would be less confusing to identify batches by a letter, A through H, instead of a number.

be put in small batches of their own and adjudicated manually, again using the same interface. No ballots required this special handling in this pilot audit.[20]

During the scanning process, an election official manually created an Excel ballot manifest that enumerated the batches (such as P1-4) and the number of ballots in each batch *as annotated by the election officers who created the batches*. Those batch counts then were compared to the counts reported by the scanner (on its display) and by the audit software, for each batch and aggregated for each precinct. This ballot manifest defined the sampling "universe" from which ballots would be randomly sampled.[21] If any of the counts had disagreed, the election officials would have investigated the cause of the discrepancy and, if necessary, rescanned the batch or corrected the ballot manifest. The image here shows part of the ballot manifest used in creating the sample: a comma-separated value (CSV) file as rendered in Excel.[22]



	A	B	C	D
1	precinct	batch num	batch ID	batch size
2	CAP	1	CAP-1	25
3	CAP	2	CAP-2	25
4	CAP	3	CAP-3	11
5	P1	1	P1-1	25
6	P1	2	P1-2	25

Once all the ballots had been scanned and retabulated, and the results compared with the voting system results, it was time to select the random seed and generate the random samples.[23] The team decided to use an approach similar to that used in Colorado's risk-limiting audits in November 2017 and July 2018, using twenty ten-sided dice of different colors, similar to those in the picture to the right. The names of observers were placed in a hat; the dice were placed in a second hat. Observers around the room took turns choosing a name from the hat; each person selected then would blindly draw a die and roll it in front of the other observers. The

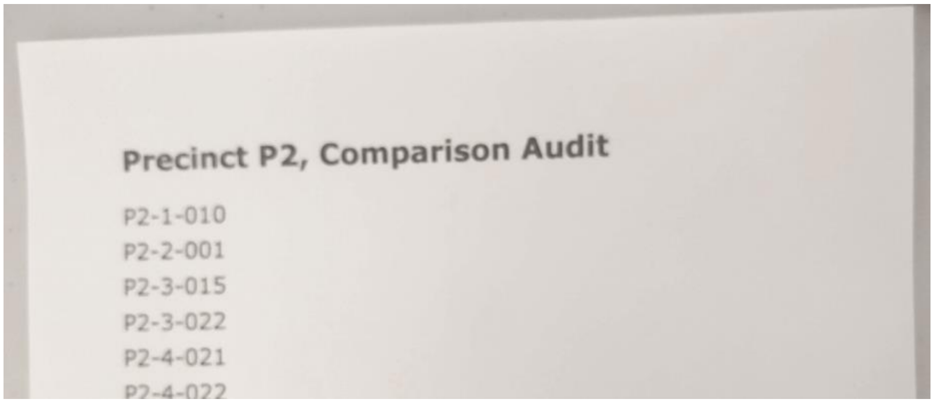
[20] No ballots were cast using the accessible ballot marking system, the Unisyn OVI, which produces a selections-only ballot about 3¼ inches wide. Any such ballots also could have been adjudicated manually, although doing so could violate voter privacy if a selections-only ballot could be traced back to the voter. Audits and recounts often foreground such privacy concerns

[21] That is, even in a "pure" ballot-polling audit in which Cast Vote Records were not available, we can draw a random sample from across all the ballots by knowing how many ballots are in each batch. For instance, batch P1-4 contained 25 ballots, so it is known to contain ballots P1-4-001 through P1-4-025.

[22] This manifest, strictly, is not the version manually created by election officials, but a version created by the audit software and validated against the manual version. See footnote 31.

[23] At this point, the protocol omitted a useful step. Properly, the retabulation Cast Vote Records should have been "hashed" – that is, a quasi-unique, tamper-evident digital identifier should have been computed from the data file – and at least the hash value should have been distributed to observers before the random sample was generated. The Cast Vote Record file could be published at that time or later, allowing observers to check that the file was unchanged (because it produces the correct hash) and that the comparisons were done correctly. Instead, the Cast Vote Record file was distributed to election officials after the random sample was generated. This omission was a concession to limited internet connectivity and the modest goals of the pilot. It did create a subtle gap in audit observability: in principle, the audit system could have been compromised to "swap" correct and incorrect Cast Vote Records so that only correct CVRs would appear in the audit sample.

resulting twenty-digit seed, 78273138432832015441, is plausibly random.[24] This random seed, the ballot manifest, and the desired sample sizes then were provided to a sampling function closely based on a publicly available Python function written by MIT computer science professor Ron Rivest, using a well-known pseudo-random number generator (PRNG) called "SHA-256 in counter mode."[25] This implementation ensures that the ballot sample (1) is completely unpredictable before the seed is selected, and (2) can be replicated by anyone with access to the software once the seed is selected. This approach supports the principle of observability: it enables observers to verify the validity of the audit in detail. The ballot IDs of the resulting samples were printed on sheets by precinct and sample (ballot-level comparison or ballot polling) to facilitate retrieval. Part of one of these "pull sheets" is shown below.



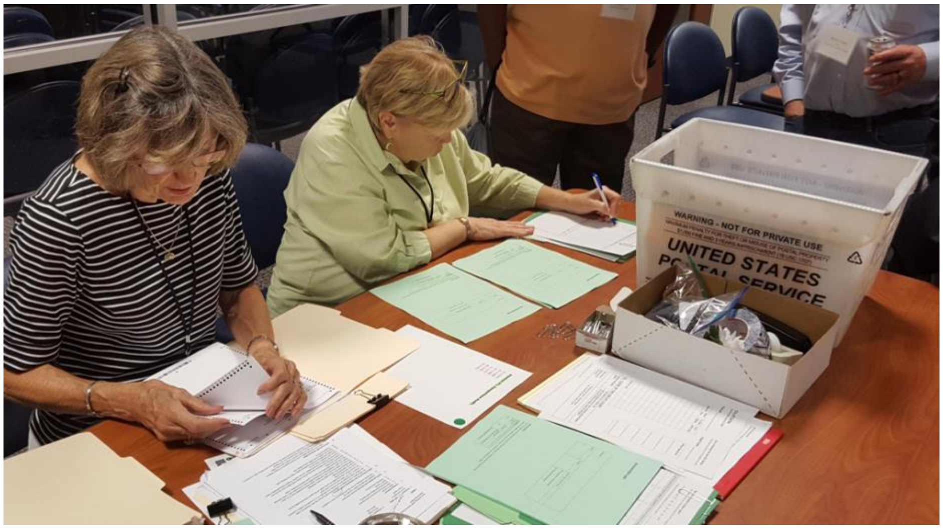
Precinct P2, Comparison Audit

P2-1-010
P2-2-001
P2-3-015
P2-3-022
P2-4-021
P2-4-022

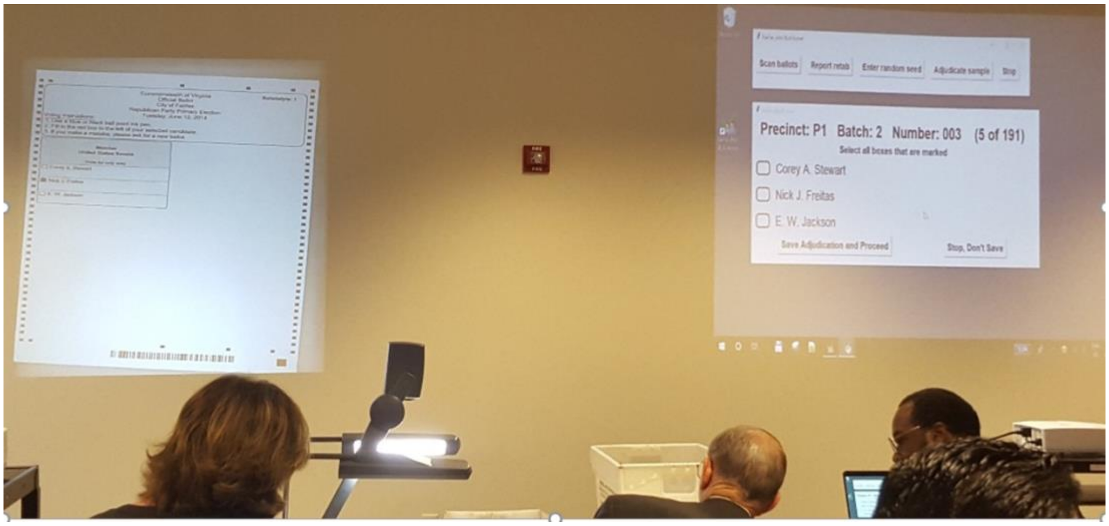
On Friday, the teams of election officers retrieved the ballots in the two samples – first the ballot-level comparison sample (so adjudication of this smaller sample could begin as soon as possible), then the ballot-polling sample. For each retrieved ballot, the teams used one sheet of colored paper – light green for comparison, yellow for ballot-polling – manually labeled with the appropriate ballot ID, to hold the ballot's place in the batch, and a second identical sheet to identify the ballot itself as it went to the adjudicators. Thus, the adjudicators were given interleaved stacks of colored cover pages and voted ballots. (The team considered using the audit software to produce these placeholder and cover pages, but ruled it out because of logistical complications – no high-speed printer was readily available – as well as limited development time.) The photo below shows one election judge counting through a stack of voted ballots to retrieve the next ballot on the pull list, while the other fills out a ballot tracking worksheet. Notice the green cover sheets, the list of procedures (white sheet at the bottom of the picture), and the box of supplies.

[24] Although observers found this approach highly credible, it might be improved upon. Neal McBurnett has pointed out that under certain circumstances, an observer might be in a position to subvert the audit by computing all possible audit samples conditional on the first eighteen or nineteen digits, determining which sample(s) (if any) would falsely confirm the voting system outcome, and then somehow ensuring that the last one or two dice were loaded to produce the desired sample.

[25] The reference code can be found at https://people.csail.mit.edu/rivest/sampler.py. The version used in Fairfax is functionally identical, but updated for Python 3.



The adjudication process involved two adjudicators and a software manager, all of whom were local or state election officials. For each ballot, one adjudicator first read off the ballot ID on the cover page, and the other people confirmed that the software displayed the corresponding ballot ID. Then the adjudicator used an analog document projector provided by the county court to project the ballot on the wall, and one or both adjudicators verbally read off the vote(s) on the ballot, if any. The software manager entered these vote(s) in the software, and clicked to proceed. The software then displayed a confirmation dialog: “Record ballot [ID] as [Value],” where the value could be “Stewart,” “Freitas,” “Jackson,” “undervote (blank vote),” or “overvote (multiple candidates marked).” Then the adjudicators removed the ballot from the projector and examined the ballot directly; the software manager read back the vote(s) reported by the adjudicators according to the dialog, waited for assent or dissent, and either confirmed or amended the adjudication. The photo below shows the ballot P1-2-003 about to be adjudicated.



When all ballots in a round had been adjudicated, a crude text report of the results was generated. Mark Lindeman presented these results, and Brenda Cabrera filled in logistical details. During the retabulation and adjudication, several guests made presentations; observers could choose whether to watch the audit, pay attention to the presentations, or both.

Software design

Because no known open-source software had all the functions needed to support the pilot design, and because the election was simple, the planning team decided to custom-build audit software to manage the scanning, interpret and tabulate the votes, record the random seed and generate samples, store the audit interpretations of each ballot, and produce results reports. Mark Lindeman wrote the software in Python, using the Tkinter toolkit to design the user interfaces.

A design objective was for election officials to control the software to the greatest possible extent. At the same time, given the compressed development schedule and the variety of interface elements involved, it was considered important to allow Lindeman to intervene if necessary, without impinging upon the role of election officials. The team had planned to run the audit software on a laptop provided by the state, but due to permissions problems, ultimately one of Lindeman's laptops was used. State officials operated the software throughout the audit, except that Lindeman assisted in producing the lists of ballots for retrieval and the results reports, and intervened when a bug in the auditing software appeared late in the scanning process.[26]

Another design objective was to represent most data and outputs, except for actual ballot images, in easily readable .CSV (comma-separated value) and text files. (These files are listed and briefly described in the appendix.) The ballot images were stored using file names that incorporated their ballot IDs. Although this legibility did not make much practical difference during the pilot, it was consistent with the principle of observability, and it provided a means whereby election officials and observers could examine the audit results without relying on intermediary software beyond, possibly, an image viewer.

Retabulation by the software designed for the audit worked as follows: Each scanned ballot image was initially captured in 8-bit RGB color. A grayscale copy of the image was saved for reference. Then the color image was processed through a red-drop filter (see footnote 27), which removed the marking targets as well as the grain of the paper; the resulting image was stored as a compressed black-and-white TIFF image (about 60 kilobytes per image). This black-and-white image was registered against a black-and-white master image – in effect, straightening and regularizing the image. (The registration process was tuned to take about 1.5 seconds per ballot. Much faster approaches were

[26] Figuratively, after election officials stopped and restarted the scanning process, the software became confused about which precinct it was scanning. (The previous day's testing did not detect this bug because the scanning process was never interrupted.) It took about five minutes to correct the code and manually correct some ballot IDs.

possible, but this process was designed not to depend upon timing marks.[27]) The software then used its knowledge of the master image to locate each of the three voting targets in the registered image and measure the darkness of each target – that is, the proportion of the pixels that were darkened. Any mark that was more than 10% dark was coded as a vote attempt; and each ballot was coded as a vote, undervote (no vote attempts), or overvote (multiple vote attempts) accordingly. The three darkness values as well as the adjudication were stored in software.

Perhaps the greatest shortcoming of the audit software was in not logging events to provide more detailed timing data. It was intended that election officials would be able to time various processes as they occurred, and ELECT officials did do so with some success – but often they were distracted by events. An event log not only would document the use of the audit software itself, but it would help in indirectly measuring some of the "upstream" delays in batching or retrieving ballots.

Procedural comments

Overall, the pilot went remarkably smoothly, albeit not quite as well as more familiar election processes sometimes run. At a few points, election officers became confused about how many ballots they had counted (to place in a batch or to retrieve a specific ballot), which batch they were supposed to be retrieving from, or which ballot was which. These small problems were to be expected given the unfamiliarity and even strangeness of dividing ballots into batches of 25. With one exception described below, officials caught and corrected all these mistakes without affecting the audit in any way. That is an impressive accomplishment.

The schedules were designed to be conservative, and by all measures they were. On Thursday, ballot batching and scanning was projected to begin at 10:30 (after some preliminaries, including training the per diem election officials and signing out ballots) and continue until 4:30 with a one-hour lunch. Despite being manually intensive, these steps ran so far ahead of schedule that audit officials decided to extend the one-hour lunch to two hours, and to summon the state election officials not yet in attendance to come as soon as possible.[28] On Friday, retrieval and adjudication again ran hours ahead of schedule. The timing data are not directly relevant to other jurisdictions or even different use cases: we know from other audits that scanning and retrieval could be done more quickly with different equipment and processes, whereas adjudication of more complicated ballots would take longer. Nevertheless, thanks to careful advance planning by election officials, observers witnessed a process that was intelligible and manageable, not arcane and horrific. Many observers commented on how easy everything seemed.

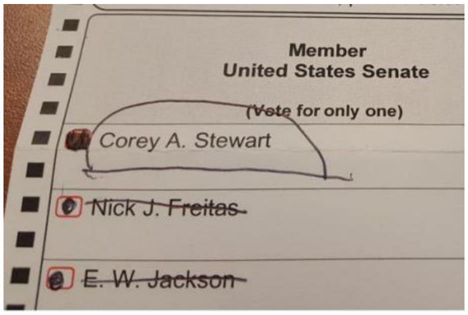
[27] The initial software design used a "sample ballot" template with a different header than the voted ballots, and without timing marks. Accordingly, the software was designed to work even if a better template never became available. (The final master template was created a few days before the pilot.)

[28] The scanning time could have been further reduced by simplifying image processing and registration and omitting manual review of undervotes in each scanned batch. However, the scanner usually stood idle: manually creating batches of 25 ballots was the main bottleneck. (Small batches also complicate the scanning process, although they may reduce the chance of jams.)

Formal results

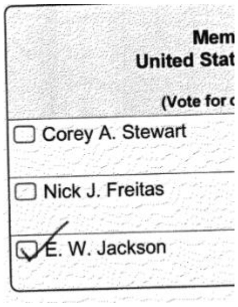
Audits often produce unexpected small insights, and the Fairfax pilot was no exception. The process of “batching” and rescanning voted ballots revealed an additional unrecorded ballot, an undervote, in Precinct 1. It is believed that this ballot had adhered to the ballot above it throughout the voting process. An election official commented that he had heard this scenario offered as a possible reason for canvassing discrepancies, but he had considered it an “urban legend” until now.

Both the hand count of the absentee ballots, and the manual review of undervotes and overvotes, directed attention to an absentee ballot that had originally been counted as an overvote (see image, right). Election officers and adjudicators unanimously agreed that the ballot would be counted as a vote for Corey Stewart under Virginia’s guidelines for determining voter intent during hand counts. Some election systems support automatic review of overvotes and marginal marks before results are announced.



With these two exceptions, the rescan retabulation matched the original voting system results in each precinct. The hand count of the Central Absentee Precinct likewise matched the original results except for the overvote pictured above.

Although many voters did not fully fill in the targets as instructed, the audit process and subsequent review of the data found no cases in which marginal marks, with darknesses close to the threshold, created ambiguity about how a ballot should be counted. The closest case was ballot P5-4-013; a portion of the scanned image, prior to dropping the targets, is reproduced here in grayscale.[29] The software reported that the target for Jackson was just 13% darkened, but looking at the entire mark, this appears to be an unambiguous vote for Jackson under Virginia hand count guidelines. Given that the retabulation vote count for precinct 5 matches the original scanner count, the Unisyn voting system evidently agreed. Six additional ballots had targets between 20% and 25% darkened, but subsequent review shows that they all contain unambiguous check or X marks for a candidate.



The results of the ballot-level comparison (70 ballots, 69 unique ballots) produced a surprising discrepancy: the report indicated that one vote tabulated for Freitas had been adjudicated as a vote for Jackson, and vice versa. Officials quickly established that the

[29] Notice how the grayscale image shows the grain of the paper; the red-drop filter largely removed such artifacts. The filter simply isolated the red channel of the color image, whitened pixels whose “redness” exceeded a threshold of 170 (2/3 of the maximum value, 255), and then converted the result to black-and-white. In RGB color space, white is represented as 255/255/255 – a mixture of pure red, green, and blue light – and a light gray might be 200/200/200 or higher.

ballots in question were two ballots retrieved from the same batch. The election officers involved told Brenda Cabrera that at one point they had become confused about which cover page belonged to which ballot. All evidence, including the scanned images, pointed to the conclusion that they had guessed wrong. This small mistake underscores the value of being able to imprint IDs in ballot-level comparison, and arguably also the value of extended practice with audit processes. At the same time, it illustrates that RLAs can be fault-tolerant when small errors occur, regardless of their cause: the 5% risk limit was attained, with a measured risk of about 3%. It also shows how careful design can facilitate further investigation to understand the causes of discrepancies.

An additional 191 unique ballots were retrieved for the ballot-polling audit, for a total of 260 unique ballots among the sample of 300.[30] These ballots were retrieved by ballot ID, so we can consider them as an additional ballot-level comparison sample. Remarkably, this larger sample contained no additional discrepancies, consistent with the possibility that no other ballots were misretrieved or mislabeled. (It is also possible that one or more additional errors occurred, but did not produce discrepant votes.) The cumulative measured risk for the full sample – the smallest risk limit that could have been attained from these results – was less than 0.01%.

The ballot-polling audit itself, not unexpectedly, did not attain or even approach the risk limit. The sample contained 128 votes for Stewart, 108 for Freitas, and 61 for Jackson. This result was broadly consistent with the announced vote (a 6.7-point margin is not far off from the reported 11-point margin, taking the sample size into account), but the measured risk was 47%. This unimpressive result provided an opportunity to remind observers that risk does not measure "the chance that the election outcome was wrong," but rather the strength of the evidence from this audit sample (treated as a ballot-polling sample) that the outcome was *right* – the lower the risk, the stronger the evidence. A sample margin of 128 to 108, by itself, is not strong evidence. In an actual ballot-polling audit – presumably of a larger election – the sample would be expanded to collect more evidence.

Comparison with the "Principles and Best Practices"

The 2018 "Principles and Best Practices for Post-Election Tabulation Audits"[31] enunciates nine principles that tabulation audits should follow. We can consider both how well the pilot conformed to these principles and how well it supported future progress toward adherence. Overall conformity was excellent, with the crucial and inevitable exception of principle 8.

[30] The ballot samples were drawn with replacement, and some ballots were sampled twice; one ballot was sampled three times! This is unsurprising given that the sample size, 300, is indeed a considerable fraction of the 948 ballots cast.

[31] https://www.verifiedvoting.org/wp-content/uploads/2018/11/Principles_Best-Practices_Tabulation-Audits-20181023.pdf

1. **Examination of voter-verifiable paper ballots:** Hand-marked paper ballots were used throughout the audit.

2. **Transparency:** Members of the general public were not invited to the pilot, but the event modeled many best practices of observability. Observers were able to watch at a short distance as election officials scanned, retrieved, and adjudicated ballots. Although the data that would have allowed the public to check the audit results was not published, this could readily be done in the future.[32]

3. **Separation of responsibilities:** In most local pilot audits, local election officials do have discretion over how the audit is conducted, technically violating this principle. In the Fairfax pilot, unusually, state and local officials collaborated in designing and documenting audit procedures. This process did provide a check upon local discretion; more importantly, it helped the state prepare to regulate future audits.

4. **Ballot protection:** The chain-of-custody procedures we observed appeared sound. We did not inquire into details of how ballots are stored at the courthouse, but ballot containers were sealed at the beginning of the audit and were resealed at the end. The ballot counts were repeatedly checked against election night reports.

5. **Comprehensiveness:** All voted ballots were subject to being audited. Only one contest was on the ballot, so the pilot set no meaningful precedent for comprehensiveness in contest selection.

6. **Appropriate statistical design:** Given the premise that "Republican Senate primary winner in the City of Fairfax" is a meaningful outcome, the statistical design of the ballot samples yielded a valid risk-limiting ballot-level comparison audit, and a valid risk-measuring ballot-polling audit. (Hand-counting a single, preselected precinct was *ad hoc*.)

7. **Responsiveness to particular circumstances:** The pilot offered no explicit mechanism for additional auditing – and the audit was very intensive as it was. Future regulations should allow particular precincts or ballots to be audited if circumstances warrant.

8. **Binding on official outcomes:** As we have seen, it is currently impossible for any Virginia audit to alter official outcomes or results. Progress on this crucial principle will require new legislation.

[32] In Colorado, cast vote records are not yet published because of concerns about inadvertently revealing how specific voters voted, due to some extremely rare ballot styles. To our knowledge in Virginia, and certainly in the City of Fairfax, publishing cast vote records should not jeopardize voter privacy, provided that ballots printed by a ballot-marking device are not identified as such and that the records cannot be matched to information about the order in which voters cast their ballots. Manually recording the audit adjudications would allow observers to check them against published cast vote records.

9. **Investigating discrepancies and promoting continuous improvement:** Election officials closely scrutinized the few discrepancies that emerged, and they noted possible future process improvements.

Ten key takeaways

1. Focused collaboration can be highly productive. Local officials, state officials, and outside advisors established effective collaborative relationships – each doing their own work while coordinating through conference calls and email messages. Clear goals and distinct responsibilities allowed rapid progress.

2. Limited pilots can facilitate collaboration. When every decision is subject to change in subsequent audits, it is easier to reach consensus for the short term. And it was far simpler for state officials to deal with one local jurisdiction than to manage dozens of circumstances and relationships.

3. Take time to specify and document details. Despite its small size, the pilot audit depended on many operational details, in part because of the need to manually "rebatch" all the ballots. Carefully planning and documenting these details not only enabled the participants -- including per diem election officers -- to implement the audit with minimal confusion, but will help other jurisdictions to implement their own audits.

4. Leave room for local variations. Although the specific procedures used in the Fairfax pilot provide an exemplar, legislators and officials should be wary of writing details into statute or regulation. For instance, manually rebatching ballots into batches of 25, and then manually counting to locate a specified ballot, would be prohibitively difficult in a large election – but it worked remarkably well in this one. In Colorado, a deliberative process involving state and local officials and audit specialists helped in determining what procedures and variants to specify in rule, and what to leave to local discretion.

5. Design audits for observability. Preparing for dozens of observers undeniably complicated several aspects of the planning, but it reaped tangible benefits. Although not all aspects of the audit were fully observable (see footnote 23), in general people in attendance could watch all steps of the audit, see in detail how ballots were retrieved and adjudicated, and satisfy themselves that the audit was being correctly conducted. Overhead and digital projectors greatly benefited observers.

6. New voting systems should be selected with auditability in mind. Ballot-level comparison has considerable appeal, but the appeal is blunted if all the ballots must be rescanned first! Virginia's next generation of precinct-count scanners could incorporate new features that facilitate auditing – and these new systems could be adopted incrementally.

7. …but even modest capital investments can improve audits. In the Fairfax pilot, the key hardware was an $850 commercial scanner. A similar or slightly better scanner, with enhanced open-source software, likely could retabulate all the ballots in any of 60 or more

Virginia cities and counties in a day or so (often much less). No matter what one thinks of that prospect, it illustrates the point. For ballot-polling audits, a $250 counting scale can help retrieve specific ballots – and so on.

8. Laws and procedures, too, should be designed with audits in mind. Ultimately, to implement true risk-limiting audits will require audits to be completed *before* election outcomes are final. That objective might entail changes to several parts of Virginia election law, not just the audit law. For instance, it could be appropriate to extend the election calendar or to allow ELECT to adjust some deadlines for cause. Other laws and procedures may have to change, for instance, to facilitate creating ballot manifests before an audit. We make no specific recommendations for Virginia here; our point is general.

9. RLA pilots are not about risk limits. Risk-limiting audits, first and foremost, are tabulation audits – manual inspections of selected paper ballots. What observers saw in Fairfax was a well-organized tabulation audit that illuminated the process and produced real evidence about the accuracy of the machine count. The risk calculations for the hypothetical contest, “winner of Fairfax,” rightly did not preoccupy observers. Although designing to attain a risk limit can be instructive, pilots should not prioritize attaining contrived risk limits over other objectives.

10. …and even routine tabulation audits are not *all* about risk limits. Audits that do not attain risk limits can provide valuable information about tabulation accuracy. Contrariwise, it is possible to design nominally risk-limiting audits that provide very little information. Verified Voting favors the rapid adoption of risk-limiting audits in federal elections – and in other elections, to the extent feasible – but audit policy should look beyond the “top of the ticket” and take a broader view. In particular, local election officials in Virginia should be encouraged and assisted in implementing audits of local contests, even if these audits are not risk-limiting.

Conclusions

The City of Fairfax pilot made a strong case for the feasibility of future risk-limiting audits in Virginia. General registrar Brenda Cabrera spoke powerfully about what she called her “RLA journey”: beginning as a skeptic who intended to wait and see what the state would eventually require, and then emerging as a supporter of RLAs who looked forward to promoting them and, if possible, incorporating them as standard practice. Other local officials expressed interest in conducting RLAs themselves. From all accounts, the pilot – together with Cabrera’s subsequent efforts to inform her peers – have shifted the prevailing mood about RLAs among Virginia local election officials from hostility toward open-mindedness and even interest.

Virginia still has considerable work to do before it can fully and efficiently implement risk-limiting audits in future elections. Colorado passed its original statute to require statewide risk-limiting audits in 2009; its first RLA of a statewide contest comes in November 2018. In between, Colorado

- conducted multiple pilots
- adopted new voting system standards and rolled out new systems statewide
- commissioned custom support software
- convened a “representative group” including state and local election officials and audit specialists around the country
- wrote and repeatedly revised a comprehensive set of rules for the audit and related procedures

The lesson is *not* that it takes nine years to implement statewide risk-limiting audits. Colorado could have adopted RLAs in some form much sooner – and Virginia’s path in many ways is easier than Colorado’s. In 2009, many Colorado voters still used Direct Recording Electronic systems that either produced “Voter-Verifiable Paper Audit Trails” that were horrific to audit, or produced no voter-verifiable record at all. Virginia in 2020 will not have audits as good as Colorado’s, but it can be far ahead of where Colorado was in 2011!

That said, Colorado’s experience does show how taking a patient, incremental, yet persistent approach to audits can yield tremendous returns over time. Although Virginia’s voting systems have little in common with Colorado’s, much of Colorado’s policy-making approach can be emulated in Virginia and around the country.

Verified Voting could not be happier about the pilot outcomes and the outlook for verifiable elections in Virginia. We look forward to continuing to collaborate with Virginia election officials in the future.

Appendix 1: Schedule of Fairfax pilot audit

Thursday, August 2		
9:00-9:15 AM	Introduction	Brenda Cabrera, General Registrar, City of Fairfax
9:15-10:00 AM	Election Officer Training and Swearing In	Electoral Board, City of Fairfax: Curt Chandler, Rick Herrington, Lorraine Koury Brenda Cabrera
10:00-10:30 AM	Signing Out Ballots	Rowdy Batchelor, Civil Case Records Manager
10:30 AM-12:00 PM	Ballot Preparation and Scanning; Batch Comparison Audit	Election Officers: Pam Cunningham, Dennis Egan, Jo Ann Gundry, James Roberts, Susan Sladek, Beth Toth
12:00-1:00 PM	Lunch	On your own
1:00-4:30 PM	Continue Ballot Preparation and Scanning Create Ballot Manifest	Election Officers Eugene Burton, Voting Technology Coordinator, Department of Elections
4:30-5:30 PM	Random Ballot Selection	Chris Piper, Commissioner, Department of Elections
5:30-6:00 PM	Preparing Ballot List for Ballot Retrieval	Election Officials Eugene Burton
Friday, August 3		
9:00-9:15 AM	Introduction	Chris Piper David Meyer, Mayor of Fairfax
9:15-9:35 AM	Signing Out Ballots	Rowdy Batchelor
9:35 AM - completion	Ballot Retrieval	Election Officers
10:05-11:05 AM	Presentation	Jerome Lovato, Certification Program Specialist, Election Assistance Commission
11:05 AM-12:25 PM	Start Ballot Adjudication	Adjudicators: Mindy Scott, election officer; Rick Herrington Eugene Burton
12:25-1:25 PM	Lunch	On your own
1:25-2:25 PM	Results of the Ballot Comparison Audit	Mark Lindeman, Verified Voting
2:25-3:25 PM	Presentation	Monica Crane Childers, Democracy Works
3:25-4:25 PM	Q&A Session	Brenda Cabrera, Electoral Board members Dept. of Elections: Eugene Burton; Samantha Buckley, Policy Analyst Verified Voting: Mark Lindeman, Marian Schneider

4:25-5:25 PM	Results of the Ballot Polling Audit	Mark Lindeman
5:25-5:35 PM	Closing Remarks	Chris Piper

Appendix 2: results data files in order of production

This file is produced incrementally as ballots are scanned and retabulated:

- rescans.csv: CSV file containing for each ballot: precinct, ballot ID, Stewart darkness, Freitas darkness, Jackson darkness, rescan adjudication [Stewart, Freitas, Jackson, overvote, undervote]

These files are produced simultaneously when the scanning phase is concluded:

- manifest.csv: CSV file containing *software-produced* version of ballot manifest (precinct, batch number, batch ID, batch size) based on the rescans[33]
- precinct totals.txt: text file tabulating the rescan adjudications by precinct and the totals
- sample size.txt: text file containing the sample size for round 1 of the sample, computed to attain a 5% risk limit if no more than one one-vote overstatement were found (this value could be overridden by editing the file)

These files are produced simultaneously when a 20-digit random seed is saved:

- seed.txt: text file containing the 20-digit random seed
- full-sample.txt: text file containing the 300 ballots in the full (two-round) sample, in random order with duplicates
- sample1.txt, sample2.txt: text files containing the two samples, each in sorted order with duplicates removed, with the size of the first round determined by sample size.txt
- comparison-pull.rtf, polling-pull.rtf: Rich Text Format files containing the ballots to retrieve from each precinct for each round, one page per precinct-round

This file is produced incrementally during adjudication of retrieved ballots:

- adjud.csv: CSV file containing for each ballot: ballot ID, vote for Stewart (0/1), vote for Freitas, vote for Jackson, audit adjudication [same five categories as rescans.csv]

These files are produced after, respectively, the first and second audit round:

[33] This file was validated against the manually prepared ballot manifest, and then used to draw the samples. Using the software copy allowed state election officials to alter the formatting of the manual version without fear of affecting the results – but this complication would have been avoided if planning time had permitted.

- report1.txt, report2.txt: text files containing key summary statistics for the audit based on all ballots adjudicated so far (properly accounting for duplicates), including P values for ballot-polling and ballot-level comparison audits

Appendix 3: more information on basic RLA methods

Other references give more details on the basic sampling methods generally used in risk-limiting audits, as well as some of the operational considerations.[34] Here we provide a brief sketch.

Ballot-level comparison: The most efficient auditing method – the one that typically can confirm a correct tabulation outcome while auditing the fewest ballots – is called ballot-level comparison. In ballot-level comparison, individual ballots are sampled, the vote(s) on each ballot are manually interpreted, and the audit interpretation of each ballot is compared with the corresponding voting system interpretation. Voting system interpretations generally are recorded in digital Cast Vote Records, one record per ballot. (These records commonly are represented as rows in a spreadsheet table.) When feasible, ballot-level comparison can efficiently examine a representative sample from across an entire election. It is far more informative to audit a random sample of 500 *ballots* cast in various places, and find out whether each one was interpreted and counted correctly, than to hand-count 500 ballots from one precinct.

Using ballot-level comparison, many outcomes can be confirmed at a small risk limit by auditing fewer than 100 ballots. A reasonable sample size estimate for a 5% risk limit (with some error tolerance) is 7.6/m ballots, where m is the proportional margin. Thus a 10-point margin might entail auditing around 70-80 ballots; a 5-point margin might entail auditing about 150 ballots; a 1-point margin might entail auditing about 760. These sample sizes apply in all elections except very small ones: for instance, a statewide contest with a 5-point margin requires about the same amount of auditing as a congressional contest with the same margin.

However, most current U.S. voting systems do not support ballot-level comparison audits. Many systems do not save Cast Vote Records; others provide no way to match ballots with their Cast Vote Records. This limitation helps to protect voter privacy in situations where voters could be linked to their ballots – for instance, in voting locations where the order in which voters vote (or even sign in to vote) is recorded and preserved. Central-count systems used to count absentee and mail ballots – and, in some jurisdictions,

[34] A brief but somewhat technical introduction is Mark Lindeman and Philip B. Stark, "A Gentle Introduction to Risk-limiting Audits," *IEEE Security and Privacy* Special Issue on Electronic Voting, 2012, available at https://www.stat.berkeley.edu/~stark/Preprints/gentle12.pdf. A broader discussion can be found in a white paper by the Risk-Limiting Audits Working Group: "Risk-Limiting Post-Election Audits: Why and How," October 2012 (version 1.1), available at https://www.stat.berkeley.edu/~stark/Preprints/RLAwhitepaper12.pdf. New methods have been developed since these papers were published, and additional methods are in development.

ballots originally cast at a polling place – do often support ballot-level comparison audits, sometimes after small modifications. Future precinct-count voting systems likely will provide ways to link ballots to Cast Vote Records without compromising voter privacy.

In the meantime, the only way to conduct a ballot-level comparison audit in many places is a transitive audit or machine-assisted audit. A transitive audit entails rescanning and retabulating all the ballots in order to produce Cast Vote Records that can be matched to ballots. If the outcome of the retabulation matches the voting system outcome – the vote counts do not have to match exactly – then a ballot-level comparison audit that confirms the retabulation outcome also confirms the (identical) voting system outcome. Although the retabulation results need not match the original vote counts, comparing these results can yield additional insight into voting system performance.

Transitive audits should not be confused with so-called "image-based audits" that rely upon software analysis of ballot images captured by the voting system or a retabulation system. Image analysis can be very useful, but the *sine qua non* of valid audits is to inspect actual paper ballots.[35]

Ballot polling: Ballot-polling audits are like ballot-level comparison audits in that they entail auditing a random sample of individual ballots. However, unlike ballot-level comparison audits, the ballots are not compared to voting system interpretations. Instead, they are treated much like responses in public opinion polls (hence the name): a sufficiently large preponderance of votes for the reported winner can provide strong evidence that the tabulation outcome is correct. This approach is considerably less efficient than ballot-level comparison, and the sample sizes needed to attain the risk limit increase rapidly for small margins. However, where ballot-level comparison is impossible, ballot polling can provide a feasible alternative that requires only the ability to select a random sample from among the voted ballots.

A reasonable estimate for the *average* sample size needed for a ballot-polling audit to reach a 5% risk limit is $6/m^2$. Thus, a 10-point margin might entail, on average, auditing about $6 / (0.1)^2 = 600$ ballots; a 5-point margin, about 2,400 ballots; a 1-point margin, about 60,000 ballots. Again, these sample sizes do not much depend on the size of the election except when the sampling fraction – the sample size as a fraction or proportion of total ballots cast – is large.

Depending on random sampling error – luck of the draw – some ballot polling samples have to be substantially larger than this average to attain the risk limit, even when the original counts are accurate. Especially for a multi-jurisdictional audit, a conservatively large initial sample size has advantages: it reduces the chance of having to coordinate additional auditing across jurisdictions, and it produces stronger evidence about the actual vote shares.

[35] For further discussion of this topic, see for instance, Mark Lindeman, Ronald L. Rivest, and Philip B. Stark, "Machine Retabulation is not Auditing" (24 March 2013), https://www.stat.berkeley.edu/~stark/Preprints/retabNotAudit13.pdf

Even when a ballot-polling audit sample is divided across many jurisdictions, individually retrieving large numbers of ballots can be unwieldy. A variation called Bernoulli ballot polling can ease some implementation challenges: it allows auditing to begin before all the votes have been counted, even on election night. Several methods are available for retrieving the desired ballots, including imprinted IDs, manual counting, and the use of specialized counting scales.

Batch-level comparison: The most common method for conducting post-election audits (but not risk-limiting audits) is generically called batch-level comparison. A "batch" refers to a group of voted ballots for which vote counts for each candidate are available, and which are stored together. For instance, all the ballots cast in one precinct on election day might constitute a batch. Many states have audit laws that require randomly sampling some percentage of precincts, or of voting machines, and comparing hand counts of the associated batches with the voting system counts.

Batch-level comparison typically requires auditing many more ballots than other RLA methods, because hand-counting a batch – no matter how many ballots it contains – provides essentially no information on how accurately any other batch was counted. Hand counts, especially of large batches, can be intricate and error-prone, although most errors are small. That said, batch-level comparison is conceptually straightforward and, in comparison to ballot polling, provides more intelligible results for local jurisdictions participating in a statewide (or other coordinated) audit.

(To see this point, imagine a local jurisdiction either hand-counting one precinct in a batch-level comparison audit, or retrieving and auditing several ballots in a ballot-polling audit. The hand count will probably take longer, but it will allow the jurisdiction to report a specific result for that precinct: "the hand count matched the machine count," or "there was an X-vote discrepancy probably due to…." Reporting the local portion of a ballot-polling audit – perhaps that "we found 2 votes for [the reported winner] and 2 votes for [the reported loser]" – is less informative, although this result constructively contributes to a valid audit.)

The sample size *in batches* of a batch-level comparison audit at a 5% risk limit can be crudely estimated as 3.2/*m*: about 32 batches for a 10-point margin, 64 for a 5-point margin, or 320 for a 1-point margin. (Again, this number does not much depend on the size of the election except when the sampling fraction is large – which is more common than with ballot-based methods.) This estimate assumes that batches are sampled not in a simple random sample, in which every batch is equally likely to be sampled, but with "probability proportional to error bound": in general, large batches are more likely to be sampled.[36] So the average batch in the sample generally is larger than the average batch in the election – perhaps much larger, depending on how variable the batches are.

This sensitivity to batch size can complicate RLAs. In Virginia, many jurisdictions have large numbers of absentee ballots – over 50,000 cast in 2016 in Fairfax County (not the

[36] The error bound depends not only on the number of ballots cast, but on the vote counts: for instance, the more votes are reported for the reported winner, the larger the error bound.

City of Fairfax) alone – that, by default, are reported as very large batches. These large batches are disproportionately likely to be included in the audit sample, vastly increasing the workload. One strategy to address this problem is to subdivide large batches into smaller batches for which totals can be reported. This can involve either sorting the ballots by precinct, or obtaining subtotals for arbitrary batches of ballots (a function that many election management systems do not currently support) and revising the ballot manifest accordingly. Another strategy is to audit these ballots using one of the other methods: ballot-level comparison (which generally entails imprinting identifiers) or, less efficiently but perhaps easier to implement, ballot polling.[37]

Hybrid audits: As the preceding sentence underscores, it is possible to use different audit methods for various ballot types or jurisdictions, combining the results to produce a valid risk-limiting audit. This ability to "mix and match" can yield advantages in many elections. Colorado could use a hybrid method to audit statewide contests, combining ballot-level comparison results in most counties with ballot polling results from the few counties that cannot match ballots to their cast vote records. In December 2018, three Michigan municipalities are conducting RLA pilots that combine ballot polling of precinct-cast ballots with ballot-level comparison (based on retabulation) of absentee ballots. Details of these methods, and other ways in which RLAs can be adapted to circumstances, are out of scope for this report.

[37] One statistical approach to combining comparison and ballot-polling samples is described in Kellie Ottoboni, Philip B. Stark, Mark Lindeman, and Neal McBurnett, "Risk-Limiting Audits by Stratified Union-Intersection Tests of Elections (SUITE)," available at https://arxiv.org/abs/1809.04235.