Risk-Limiting Audit Recommendations for Connecticut

RLA Recommendation Working Group

February 2, 2022

1 Executive Summary

The Risk-Limiting Audit Working Group was established by Public Act-21-2, June Special Session, to study the benefits and drawbacks associated with implementing risk-limiting election audits in the State of Connecticut and formulate a recommendation.

Risk-limiting election audits. A risk-limiting audit (or RLA) is a post-election audit [1–4] that evaluates, by hand, a subset of cast ballots in order to detect circumstances where the winner reported by the tabulation disagrees with the winner indicated by the physical ballots. An ideal risk-limiting audit provides two guarantees:

- Statistical bounds on risk. The audit should provide explicit, mathematical guarantees on the probability that an invalid election is not detected. (The probability of such a failure is called the risk of the audit.)
- Transparency. The audit should be sufficiently transparent to convince an observer of correctness.

Connecticut's current audits are not risk-limiting. Connecticut's current auditing procedure hand counts—or counts with electronic assistance—5% of all precincts. While such an audit can detect even small tabulator or procedural irregularities among the selected precincts, it is not risk-limiting in the sense described above because the number of audited precincts does not scale with the margin of the results; in particular, in cases with tight races that span several precincts the audit will not provide meaningful control of risk.

Recommendation. On the basis of the study of the published literature on risk-limiting audits, risk-limiting audit pilots conducted in other states, and a risk-limiting audit pilot run in Connecticut on January 6th and January 10th, 2022, **the working group recommends that the State of Connecticut adopt comparison risk-limiting audits for its federal year elections.** Our detailed recommendation calls for the office of the Secretary of State to plan and oversee the annual audits. The recommendation is supported by a discussion of the efficiency of various methods, the operational challenges associated with the audits, the guarantees they offer, and speculation about future equipment that would simplify and streamline the procedure.

The detailed recommendation appears in Section 7 below.

2 Goal and origin of this document

The Connecticut legislature called for the establishment of a pilot program to study risk-limiting audits in Public Act-21-2, June Special Session. The relevant language from the act:

> (a) There is established a working group to (1) examine employing risk-limiting audits to determine the accuracy of election results, including (A) the feasibility of implementing such audits, (B) the different methods used in such audits and the practical considerations for implementation of each such method within the existing statutory framework, (C) any potential equipment necessary to implement one or more of such methods, (D) the procedures necessary to implement one or more of such methods, and (E) any changes to such statutory framework necessary to implement one or more of such methods, and (2) within available appropriations, oversee a pilot program in not less than five and not more than ten municipalities of one or more of such methods for the municipal elections held in such municipalities in 2021.

The established working group consisted of the following members:

1. Gabe Rosenberg, Chair, Designee of the Secretary of the State, General Counsel Connecticut Secretary of the State
2. Gemeem Davis, Appointed by the President Pro Temp of the Senate, Bridgeport Generation Now
3. Giselle Feliciano, Appointed by the Speaker of the House, Registrar of Voters, Hartford
4. Robert Ham, Appointed by the House Minority Leader, Cheshire, CT
5. State Representative Hilda Santiago, Appointed by the Chairs of the GAE Committee
6. Louis DeCilio, Appointed by the Senate Minority Leader, Registrar of Voters, Stratford
7. Dominic Rapini, Appointed by the Ranking Members GAE Committee, Branford, CT
8. Aida Carini, Appointed by the Secretary of State (Legal Expertise), Staff Attorney, Connecticut Secretary of State
9. Brian Macdonald, Appointed by the Secretary of State (Statistical Expertise), Lecturer in statistics and data science, Yale University
10. Lois Timms-Ferrara, Appointed by ROVAC, Registrar of Voters Ellington
11. Tim Beeble, Appointed by ROVAC, Registrar of Voters Bethel
12. Alexander Russell, Director Center for Voting Technology Research, University of Connecticut.

During the second meeting of the working group, a motion was passed calling for the UConn VoTeR Center to manage the pilot program and summarize its recommendations. As a result, this document is primarily authored by the University of Connecticut Center for Voting Technology Research (VoTeR Center) and has been reviewed and approved by the RLA committee.

This report describes the operational aspects of risk-limiting audits, the design of the pilot program, the working group's concluding recommendations, and justification and discussion of those recommendations.

The UConn VoTeR Center. The UConn VoTeR Center was founded in 2006 to assist the State of Connecticut to satisfy the requirements of the Help America Vote Act of 2002. The VoTeR Center works under the direction of the Connecticut Secretary of the State's office (SotS), providing technical advice and auditing tools to support Connecticut's elections.

The UConn VoTeR Center conducts comprehensive auditing of equipment used in the elections and designs safe election procedures. Starting in 2008, the Center has performed technological audits of memory cards used in Connecticut elections and assisted in tabulator audit procedures. The Center has also designed software (the "Audit Station") capable of producing cast-vote records from the ballots cast in Connecticut elections; this is a critical component of typical ballot comparison audits.

The VoTeR center began researching and developing new risk-limiting audit techniques in Spring 2021.

3 What is a Risk-Limiting Audit?

A risk-limiting audit (RLA) is a post-election audit that hand counts a subset of ballots in order to provide rigorous statistical guarantees on the correctness of the election. The term risk limit refers to the chance that the audit fails to detect a disagreement between the tabulated winner and the rightful winner determined by the physical ballots. The simplest RLA is a complete hand count of the election; this has zero risk, but requires a lot of work. This motivates risk-limiting audits based on random samples of ballots; although these are significantly less expensive, they can still provide satisfactory control of risk. Such statistical audits will be the focus of the recommendation.

In general, the number of ballots that must be hand counted during a statistical RLA depends on

- the desired risk limit, and
- the tabulated margins of the races to be audited.

Intuitively, narrower margins require more ballots to be examined because fewer errors are necessary to change who wins the election. In practice, the number of ballots that must be hand counted for a single race ranges from 30 to 30,000—the details depend heavily on the type of statistical audit and the margin of the race. In circumstances where one wishes to audit multiple races, the number of ballots that must be examined in total depends on whether the races involve the same ballots. Specifically,

- In settings where two or more races share the same ballots (e.g., during auditing of multiple statewide races), the number of ballots to be examined is approximately the number of ballots necessary for auditing the (single) race with the smallest margin among the various races. Thus, in these cases extra races can be audited more or less "for free."
- In settings where the races share no ballots, one must simply audit each race independently.

In more complicated settings—for example, settings where one race involves ballots that are a strict subset of those for another race—the extent of these efficiency improvements depends on a detailed accounting of the intersection pattern and the margins.

With a particular risk limit set in advance, two conclusions are possible at the termination of a risk-limiting audit:

- The audit has amassed sufficient statistical evidence that the tabulated outcome is correct. The probability of an error is determined by the risk limit.
- The audit is inconclusive. In this case, one could conduct further statistical audits or undertake a full recount.

Because the cost of a full hand count is so high, it is common practice to begin a subsequent statistical audit when a completed audit is inconclusive. This results in a procedure where the full audit is effectively an iterative process that proceeds in "rounds." After each round of ballots has been selected and processed, if the audit procedure does not have strong enough evidence to confirm the result then an additional round of auditing is conducted.

Risk-limiting audits can be understood as falling into four styles (see prior academic, industrial, and government work [1–14]): batch comparison, ballot polling, ballot comparison, and full hand count. Aside from the full hand count, these methods use tabulated election results to randomly determine a subset of ballots to examine.

Batch Comparison In a batch comparison audit, the ballots are assumed to be organized in "batches" with the understanding that tabulated totals exist for the individual batches. A subset of batches are randomly selected and are manually counted; the results are compared against the tabulated totals.

Ballot Polling In a ballot polling audit, a random sample of ballots are picked and hand counted. The margin of these counts is then compared to the margin of the reported results.

Ballot Comparison In contrast to the previous auditing styles, ballot comparison audits require a cast-vote record (CVR). A cast-vote record is a comprehensive declaration of the votes appearing on each ballot of the election. Typically, the CVR is a table, each row of which contains a ballot identifier that refers to a specific cast ballot followed by a listing of the votes appearing on the ballot. The audit proceeds by selecting a random collection of ballots which are then compared against their entries in the CVR. Although ballot comparison methods have the additional requirement of a CVR, they examine fewer ballots than polling and batch comparison for the same risk limit.

It is useful to distinguish two varieties of ballot comparison audits depending on how the CVRs are generated. If the election tabulators themselves produce CVRs, the audit is referred to as a "primary ballot comparison" audit. Otherwise, the CVRs must be generated by additional equipment in which case the audit is called a "transitive ballot comparison" audit. As the tabulators currently used in Connecticut do not produce CVRs, the only feasible ballot comparison audit is transitive. Connecticut does have access to equipment capable of generating CVRs.

Full hand count. A full hand count is costly, but achieves zero risk.

Benefits of Risk-Limiting Audits Risk-limiting audits are procedures designed to establish confidence in the results of an election. In comparison with informal approaches, they offer explicit, rigorous guarantees of risk and—when properly conducted—offer observers confidence in the conclusions.

Requirements of a Risk-Limiting Audit All types of RLAs require a voter-verified paper record of the election; this means physical ballots marked, either directly or indirectly, by voters. Such tabulation of hand-marked paper ballots is the standard in Connecticut. Aside from full hand counts, all methods require a ballot manifest which is a partition of the cast ballots into batches with human-verified sizes.

- Batch comparison audits require individually tabulated batches (such as those arising in Connecticut elections by voters casting ballots directly into tabulators at individual polling locations). During the audit, batches must be identified and located for hand counting.
- In polling and ballot comparison audits one needs to retrieve a specific ballot from a batch. As we discuss below, ballot retrieval is a challenge, especially in settings where ballots do not have individually identifiable marks (as is currently the case in Connecticut).
- Lastly, ballot comparison audits require CVRs.

3.1 A detailed discussion of RLA components and procedures

We discuss in more detail the components of RLAs and how they are combined to yield the audits described above.

Ballot manifests and batches. In order to conduct any statistical RLA, one must begin with a ballot manifest, which provides a trusted accounting of the locations of paper ballots after they have been tabulated. In particular, a ballot manifest must be created (or verified) by a human. At a minimum, a ballot manifest must identify a partition of the ballots into a family of distinguished "batches" and identify the size and location of each batch. A batch of ballots is typically stored together in a single marked storage container.

The purpose of a ballot manifest is to provide authoritative validation of the total number of ballots included in each batch and, moreover, a means for selecting batches with probability proportional to their size. Assembling a ballot manifest may involve a "ballot reconciliation process" by which ballot manifest totals and electronic tabulator totals are compared to ensure that the audit team has the correct number of ballots. Even this reconciliation process itself can be a valuable auditing technique as it exposes circumstances where the number of ballots reported by the tabulator is inconsistent with the size of the associated batch of physical ballots.

The choice of how best to partition ballots into batches depends on the details of the election itself and the particular requirements of the audit.

Cast-vote records (CVRs). Ballot comparison methods require a cast-vote record (CVR), which documents the votes appearing on each ballot cast in the election. A typical CVR is a table with one row per ballot: the row must contain both the information necessary to locate the specific ballot and the votes appearing on the ballot. A ballot comparison audit proceeds by randomly selecting ballots and comparing the human counted votes on the physical ballot against the appropriate row in the CVR. If these comparisons repeatedly discover agreement between the ballot and the CVR, they statistically support the results of the CVR and can yield a controlled risk conclusion that the winner determined by the ballots is the same as the winner indicated by the CVR.

As remarked above, in principle a CVR can either be generated by the voting tabulator itself (a "primary" CVR) or specialized audit equipment (a "transitive" CVR). However, the current tabulators deployed in Connecticut do not generate CVRs; in particular, without new voting equipment, a transitive approach is necessary. Connecticut does,

however, have existing auditing equipment that can produce CVRs directly from cast ballots. This "Audit Station" [15] has been used to assist with equipment audits in Connecticut for ten years. We remark that while tabulators do exist that produce CVRs, these products are not intended to be "voter facing" but are rather intended for centralized tabulation.

Locating ballots and ballot identification. Any ballot comparison RLA needs a means to explicitly identify ballots and reliably locate the ballot. In particular, such identifying information appears in each row of the CVR. One option for such identification is a description of a physical location, such as "ballot 31 in batch 16." Prior states' pilots show that this method of identification and retrieval is error-prone. (See Section E for further discussion.) An alternative approach is to demand that ballots have printed "serial numbers" or other such indelible identifiers—these allow auditors to recognize when they have pulled the correct ballot from a stack for examination. There are two natural approaches for ballot identifiers. The first calls for ballots to simply be printed with serial numbers when they are initially printed. This is convenient, but interferes with voter privacy. The second approach calls for serial numbers to be applied either during tabulation (by the tabulator itself) or afterwards. Our pilots used post-tabulation application of QR stickers for this purpose.

Finally, ballot polling audits also require a means for selecting ballots from batches. The requirements here are weaker, since the ballot polling always calls for a random ballot to be drawn from a particular batch. This can be achieved by selecting a specifically identified ballot, in which case the previous considerations apply, or by methods that are designed to draw a fairly chosen ballot from a physical stack.

3.2 A qualitative discussion of RLA trade-offs and hurdles to adoption in Connecticut

We summarize these considerations in the context of current Connecticut election infrastructure.

- A ballot polling audit can be carried out with an arbitrary batching of the ballots, so long as a reliable manifest determines the size of each batch. Connecticut's distributed elections naturally provide such batching according to voting precinct. The procedure further requires a reliable method to select a ballot at random from a particular batch, a nontrivial operational issue. While polling audits thus have the most favorable up-front costs, they scale poorly as a function of margin. For example, a polling audit of a single race with a 2% margin and a 5% risk limit requires hand counting over 16,000 ballots. With a 1% margin this rises to over 61,000 ballots.

- A ballot comparison audit can provide significantly more favorable scaling as a function of margin. For example, a comparison audit of a single race with a 2% margin and a 5% risk limit requires evaluating fewer than 400 ballots. However, a ballot comparison audit requires a full CVR of the election and, as discussed above, requires identifiers to be placed on ballots for identification. The Connecticut Audit Station can be used to produce such CVRs, though of course this requires a second round of processing. Thus a ballot comparison audit requires more up-front preparation work, but significantly better performance in terms of the number of evaluated ballots. We remark that in typical settings where one wishes to undertake audits of many races, the same CVRs and ballot identifiers can be used.

  CVRs pose an additional issue: one must decide what portion, if any, of the CVR is made public. Since the CVR displays all cast ballots, publishing the CVR has other consequences [2]. For example, in small batches it may be possible to make statistical inferences on voter preferences. On the other hand, public availability of CVRs is important for transparency.

- A batch comparison audit requires that the ballots are organized into batches with individual batch tabulations; individual batches are then hand counted and these hand counts are compared with the tabulations. Connecticut does separately tabulate each precinct, leading to natural batches. These batches can be quite large, so that the task of hand counting them may be formidable.

The VoTeR center is currently developing new methods to streamline both batch and ballot comparison audits in Connecticut by leveraging the existing audit station and knowledge of Connecticut election procedures.

We discuss these trade-offs in more detail later in the document. Our final recommendation (see Section 7) calls for comparison audits, as these appear to minimize the effort required by Connecticut registrars of voters, especially in circumstances where many races are simultaneously audited or margins are small.

4 Findings from other State Implementations Major Factors in RLA Efficacy

In this section we summarize the major factors that determine the efficacy of an RLA. These factors are drawn from analysis of the State of Colorado's 2020 Presidential RLA Discrepancy Report [16]. We further review findings from other states' implementations in Appendix E.

1. Procedure Complexity:
   (a) Most discrepancies are due to auditors retrieving an incorrect ballot. Even when ballots are identifiable, it is a difficult task to retrieve the correct ballot from thousands. A standard practice in RLAs is for auditors to enter the chosen ballot selections without knowing the value present in the CVR. Thus, it is not possible to use vote information to determine if the correct ballot was retrieved.
   (b) In a smaller fraction of cases, audit board or adjudication judges entered incorrect information that did not correspond with the actual voter markings on the ballot.
2. Software Complexity:
   (a) RLA software standardizes contest and choice names to prevent errors from arising due to syntactic differences between precincts/counties; when the standardization function does not work, such a mismatch can be flagged as a discrepancy. Such an error occurred in the audit of the Colorado 2020 general election [16].
   (b) If a ballot is blindly entered, individuals may write the name candidate differently, for example excluding the period on a middle initial. Such name mismatches cause errors.
3. Judging Voter Intent:
   (a) Any hand recount of ballots must interpret voter's choices. When the voter's choices are ambiguous such choices may be interpreted differently by auditors and the tabulator. Note in some cases the audit board and adjudication judges disagreed on voter intent.

We compare different types of RLAs against these three factors in Table 1.

4.1 Retrieving the correct ballot; ballot identification

The most common discrepancies on the Colorado 2020 Report arose from retrieval of an incorrect ballot [16]. The best defense against such errors are ballot identifiers so that auditors can immediately recognize whether they have the correct ballot. As discussed earlier, printing serial numbers on ballots prior to the election poses complications for voter privacy. The natural alternative is imprinting, which refers to printing an identifying mark on a ballot after (or during) tabulation. An important factor to consider is the risk that a malfunctioning imprinting process interferes with a ballot's integrity. To mitigate this, one might choose ink color, positioning, or mechanical limits on the imprinting device. For our pilots, we used stickers with QR codes as a means of imprinting. This is not a perfect solution, but does clearly distinguish marks "added during imprinting" from those present on the ballot at the time of casting.

Manual ballot retrieval without identifiers is only feasible for ballot polling, where the task at hand is to "fairly" draw a random ballot from a batch. Since the mathematical analysis of the audit requires this random choice to choose each ballot with equal probability, one needs a specific procedure to help ensure this. In this setting there are several possible methods presented by Rhode Island in their 2019 Pilot Program [17]. These methods are the scale method, the counting method, the ruler method, and the k-cut method, which are all described below. In each case, the auditing software is expected to provide explicit instructions to locate a ballot, a "ballot pull sheet." The rules for interpreting the instructions on the pull sheet depend on the details of the method.

- **Scale method**: The ballot pull sheet identifies ballots by a batch identifier and a sequence number (in the batch). Ballots are retrieved from a batch by adding ballots to a scale from the top of the batch until a weight is achieved that has been established to (nearly) correspond to the requisite number of ballots. (In principle, the pull sheet could also indicate the desired weight.) While it is reasonable to assume that the minor errors that arise in this process do not bias the sample in favor of any outcome, the true extent and effect of the errors has not yet been thoroughly studied [17].

RLA Mechanism		Procedure Complexity	Software Complexity	Judging Voter Intent
Batch Comparison	Pros	Similar procedure to full ballot recount. Requires ballots to be stored in batches (which most states already do). There is no single ballot retrieval; ballots are retrieved in batches. Can be conducted at multiple locations.	Tallying-software when conducting the audit is very similar to the software utilized when a manual hand recount is conducted. Does not require a CVR.	Provides information about the accuracy of voting machines and the reporting process.
	Cons	Requires examining more ballots than any other method. Not suitable for contests with a small number of batches.	Requires machine-readable batch tallies for each physical batch of paper ballots or requires multiple batches to be jointly audited.	Time-consuming to pinpoint discrepancies because many ballots are evaluated.
Ballot Polling	Pros	Does not require re-scanning of ballots regardless of voting machine (and does not require new voting technology). Less manual work than other audit methods when the margin is not close (roughly greater than 2 percent).	Does not require CVR or other exportable data from voting machines to conduct. Only requires a voter-verifiable paper record. There is already existing software from vendors such as VotingWorks for polling audits	N/A
	Cons	High chance that the wrong ballot is pulled; complications if ballot ordering/sorting is altered. With ballot identifiers, no way to determine if correct ballot is pulled. More selected ballots than ballot comparison methods. Identifiable ballots create new security and privacy concerns.[1]	N/A	Audit does not attempt to provide information on the cause of discrepancies or how the ballot was tabulated. Does not attempt to compare human interpretation of voter intent to machine interpretation.
Ballot Comparison (Primary)	Pros	Usually requires examining fewer ballots than other methods.	Software can help in discrepancy identification and recording voter markings.	Can identify issues with how the original voter system judges voter intent.
	Cons	Without ballot identification procedures, high chance that the wrong ballot is pulled to be examined. Requires a CVR.	Maintaining ballot security and privacy while making either unique identifiers or CVR for them (or both). Requires some form of contest standardization analysis so data can be formatted for a readable CVR (which leaves room for software error).	Compares auditor interpretation of voter intent to machine interpretation.
Ballot Comparison (Transitive)	Pros	Does not require all precincts to participate, therefore there is the potential for less ballots to be needed initially (of course, this depends on the margin of error in the election overall).	Can identify issues with how the original voter system judged voter intent.	N/A
	Cons	Requires additional equipment available in central locations or at each municipality. Without ballot identification procedures, high chance that the wrong ballot is pulled to be examined. Additional steps required in order to keep ballots organized during movement from original location to location of outside audit machine. Requires more to rebatch and rescan ballots.	Requires some form of contest standardization analysis (which leaves room for software error). Software must be able to flag discrepancies when voter intent is unclear in audited ballots (for example, they only partially filled out the ballot or dust prevents the ballot from being correctly scanned).	No mechanism to identify mismatches between original tabulator totals and transitive equipment.

Table 1: Advantages and Drawbacks of RLA methods

- **Counting method**: The ballot pull sheet identifies ballots by a batch identifier and a sequence number (in the batch). Ballots are retrieved by counting either from the top of the stack downwards or vice versa, depending on the ballot location in the stack.
- **Ruler method**: The ballot pull sheet identifies ballots by a batch identifier and a sequence number (in the batch). Ballots close to the top or bottom of the stack can be retrieved via the counting method. Other ballots are retrieved by splitting the stack at distance measurement from one side of the batch that has been established to (nearly) correspond to the requisite number of ballots. (In principle, the pull sheet could also indicate the desired distance.) As with the scale method, the retrieved ballots are almost always off from the desired ballot, and the full consequences of this has not yet been studied [17].
- **K-cut method**: Rather than a specific ballot, the pull sheet simply gives a number of ballots to retrieve from a specified batch. Each ballot is drawn by a process that "cuts" the batch six times. Specifically, the depth of each cut is determined by a random number between 1-99. The batch is then cut at approximately that percentage of ballots from the top, which is to say that the batch is rearranged so that the bottom portion of the cut is placed on top of the top portion. (For example, if the number is 33, then the cut should be about one third of the way down the stack of ballots.) After six successive cuts, the ballot on top is the ballot used in the audit. (In principle,

the pull sheet could indicate the desired cut percentages.) Experimental and simulation results show that this method can generate approximately uniform samples [17].

4.2 Software requirements

Audit administration typically involves audit management software, which may generate the random samples for the audit, archive the results, carry out comparison, and compute the statistical p-values that determine when the audit can be safely stopped. The software requirements are most complicated for a Ballot Comparison Audit which must support comparison of a selected CVR row with the hand-interpreted values from the ballot. There are several considerations regarding what information to publish about the CVR, what information should be visible to auditors, and the user interface for recording vote data across a variety of ballot styles [18]. This process may be simplified by using a transitive audit to produce the CVR as this reduces the need to communicate parts of the CVR across the state. We note that ballot comparison audits can serve to identify process gaps and issues with voting equipment by attempting to determine the cause of discrepancies after the fact. Such identification is not possible with ballot polling or batch comparison as there is no CVR. As mentioned above the audit station in Connecticut is a transitive tabulator capable of producing a CVR.

4.3 How to interpret ballots?

In any RLA method, it is crucial to provide guidelines for audit workers to differentiate ambiguous from intentional marks [19]. This can minimize disagreements on voter intent between judges. Some basic guidelines already exist in CT [20]; it may be appropriate to expand these to describe common types of ambiguous marks [21].

4.4 Further discussion of operational aspects

RLAs require complex procedures that are naturally subject to human error. Minimizing these errors is necessary to ensure a tractable and accurate audit. We discuss some relevant details.

- It is essential that there are clear written explanations of RLA procedures and attention given to training. This additionally allows for observers to have detailed information about the RLA in advance.
- The type of RLA to be conducted will partially determine how ballots are grouped into batches and stored. Container choice is important for security and for ballot preservation. (Rhode Island tested DS200 ballot bins, generic plastic bins, and metal and cardboard ballot containers [17].)
- Proper space on audit day for all the batches or ballot containers is an important consideration.
- If conducting a batch-level comparison audit, decisions should be made about batch size based on workload and efficiency. If single ballots will be examined, there should be a explicit ballot retrieval procedure.
- A set of voter intent guidelines can help prevent disagreement between members of the audit board.

5 Pilot Implementation in Connecticut

Connecticut has the advantage that several states have already piloted and deployed RLAs for their elections. We described these findings in Section 4. Thus, the goal of the Connecticut pilot was to focus on factors that are unique to Connecticut or have not been previously studied. As an example, there are several existing RLA software packages including

1. Colorado's software, which is also available here `https://github.com/democracyworks/ColoradoRLA`,
2. MIT's toolkit,
3. Neal McBurnett's toolkit,
4. Rhode Island's implementation,

5. Berkeley's implementation, and

6. Arlo by Voting Works.

Based on conversations with Mark Lindeman[2] and Phillip Stark,[3] the VoTeR Center decided to focus on Arlo, which was not available for any previous pilot programs. In addition, the VoTeR center was unable to find any reports from pilot implementations that covered multiple municipalities in a single audit. (According to Luther Weeks of CT VoTeRs Count, some counties in New York have piloted multiple municipality audits. We were unable to find public reports.) Many municipalities that have previously used RLAs have centralized high-speed tabulators that are capable of imprinting. In particular, Colorado has a long history and carries out county-based tabulation (with the entire state using mail-in voting). States without such equipment (such as Rhode Island) have focused on ballot polling and batch comparison methods that do not require reliable identification of ballots. As such the pilot program in Connecticut had the following specific goals:

1. Conduct a multiple jurisdiction audit and understand any issues arising from communication between municipalities.
2. Understand how the VoTeR Center audit station could be used to generate CVRs enabling comparison RLAs in Connecticut. Such a solution must be considered in conjunction with a method to identify ballots. In the pilot, the VoTeR Center used a QR sticker gun applied by registrars to identify ballots before creating a CVR.
3. Provide a detailed analysis of Arlo risk-limiting audit software. To the best of our knowledge, this software had never been configured or used by someone outside of Voting Works.
4. Introduce RLAs to Connecticut registrars to learn what specific procedures will be most helpful. The goal is to understand the most efficient procedure.

In addition, the VoTeR center collected timing information for all stages of the process and conducted informal interviews with registrars after the fact to understand difficulties and potential opportunities to improve the procedure. In some cases, we defer to timing collected from previous pilots when we believe they are more accurate for an RLA in practice.

5.1 Overall Design of Pilot

Based on the goals identified above, a pilot program was planned to conduct RLAs in five municipalities on January 6th and 7th, 2022. (The second day was moved to January 10th due to inclement weather.) The pilot program focused on RLA methods that do not rely on hand counting of batches: ballot polling and ballot comparison. Three towns were selected to audit a single precinct on January 6th using ballot polling, two towns were selected to conduct ballot comparison RLAs with Stratford asked to audit two precincts to ensure a multi-jurisdiction RLA was included to satisfy Goal 1 above. Since CT tabulators do not produce CVRs, the VoTeR Center audit station was used to conduct a transitive ballot comparison audit to satisfy Goal 2. The VoTeR center worked with Voting Works to configure and host the Arlo software on a disconnected (islanded) network which has not been done before. VoTeR Center personnel conducted all configuration and management of software, with communication to Voting Works on software improvements. This was towards satisfaction of Goal 3. Informal interviews were conducted with registrars after the audit to understand procedural issues towards satisfying Goal 4

Because of the desire to complete each RLA in a single day, the VoTeR center chose to conduct at most a single round of auditing with parameters that were likely to demonstrate the desired risk. In addition, precincts and races were selected to produce audits that sampled enough ballots to gain timing information and familiarity with procedures but few enough ballots to complete in a single day. Registrars were asked to provide summary information of all precincts to the VoTeR Center ahead of time to choose races and risk limits. All five audits demonstrated risk under the chosen risk limit. Audits were conducted in conference rooms at the Connecticut SotS's Office in Hartford with volunteer participation from many registrars across the state (several of whom are members of the RLA Working Group). All audits were conducted using ballots from the 2021 municipal election. An overview of the pilot design is presented in Table 2. Note that in Hebron the initial plan was to audit the Board of Selectman race but that was based on a

[2]Co-director Verified Voting.

[3]Associate Dean, University of California – Berkeley and Board Advisor Open Source Election Technology (OSET) Institute. Dr. Stark was an expert witness before the CT legislature on HB 6575 concerning RLAs on March 10, 2021.

Town	Method	Precinct	Race	Risk Limit	Estimated Ballots	Margin	Samples	Hours
Hartford	Polling	7	Board of Ed	5%	208	33%	55	1.36
Norwalk	Polling	E-1	Mayor	5%	906	28%	154	3.87
Hebron	Polling	1	Question 1	10%	3000	18%	287	7.49
Ellington	Comparison	2	First Selectman	10%	920	21%	43	3.19
Stratford	Comparison	1, 2	Mayor	5%	2570	37%	27	5.52

Table 2: Overview of pilot design. Estimated Time is based on Section 6 estimates which are pulled from Rhode Island Pilot. Specifically, for polling, time is computed $\texttt{Time} = \texttt{Ballots}/4770 + \texttt{Samples} * 86/3600$ which correspond to an estimate of 4770 ballot per hour to create a manifest and 86 seconds to poll a ballot. For comparison, time is computed as $\texttt{Time} = \texttt{Ballots} * (1/4770 + 1/1000 + 2/3600) + \texttt{Samples} * (45/3600 + 86/3600$ which correspond to 4770 ballots per hour to create manifest, 1000 ballots per hour to create CVR using audit station, 2 seconds to apply a sticker, and 45 seconds to draw and adjudicate a ballot. Measured timings differ substantively. See findings in Tables 3 and 7.

misinterpretation of the ballot as that race had no losers. After doing calculations in Arlo many of the races would have required above $1,000$ ballots for a reasonable risk limit. In the pilot, we used Question 1 for audit as it had a reasonable margin that allowed for less than 200 ballots to be pulled.

5.2 Ballot Polling Pilot (January 6)

Three towns (Hartford, Norwalk, and Hebron) participated in the January 6 polling-based Risk limited Audit on SotS premises in Hartford. They brought sealed bags with ballots and the corresponding tapes from the Accuvote terminal.

The audit started at 10 AM with each town in a separate (large) room. The audits were conducted in five distinct stages. Directions were shared with the registrars at the start of each stage. Stages varied in length with the batch and manifest creation being the longest. The five stages were:

1. Batch and manifest creation,
2. Ballot sample creation with Arlo (a randomized process),
3. Ballot sample polling,
4. Ballot sample tabulation, and
5. p-value assessment.

The remainder of this section details each stage and highlights the key findings. Note that, in general, steps 2-5 need to be repeated until the outcome of the round assessment yields enough statistical confidence (specifically, a $p-$value below the requested risk limit). Risk limits and sample sizes for the pilot were chosen to most likely end with a single round and, indeed, the required p-value was met after a single round for each audit.

5.2.1 Batch and Manifest Creation

The process consisted of manually batching the ballots in groups of size exactly 50 (small batch approach) with the last batch containing less than 50 ballots. Each batch was identified by a pre-made cover sheet (placed at the top of the batch) reporting the batch number. The use of small batches eases the polling stage as the retrieval of a specific ballot for tabulation is much easier to do in a small batch. It involves counting from the front (or back) of a batch to retrieve a specific ballot. Other methods (e.g., $k-$cutting) work for larger batches but entail the repeated manipulation of the batch and an additional mechanism for selecting and locating "cuts." One downside of small batches is that it is essential to have a large enough room to make all the batches easy to access, laying flat on tables with the cover sheet visible. The importance of a good space to work, including few interruptions or noise, was noted by several registrars.

The summarized data for manifest creation is shown in Table 4. On average (weighted by sizes), registrars progressed at the rate of **1408** ballots per hour and per registrar to produce the manifest.

Once the batching is over the number of batches as well as the size of the last batch are checked against the Accuvote OS tape and encoded in the Arlo software. Note that the numbers of ballots may differ from the tape and this indeed

Town	Hartford	Norwalk	Hebron
District	7	E-1	1
Timekeeper	Alex Russell	Laurent Michel	Ben Fuller
Estimated Ballots	208	906	2723
Requested Risk	5	5	10
% to Complete	N/A	80%	80%
Auditor 1	Giselle Feliciano	Stuart Wells	Beth Fitzgerald
Auditor 2			Louis DeCilio
Random Seed	3717384634	5042324501	9342303101
Actual Ballots	209	906	2723
Start of Manifest	10:26	10:20	10:19
End of Manifest	10:40	10:57	11:15
ballots/hour, person	896	1469	1513
Audited ballots	55	138	133
Unique ballots	49	126	126
Start of Polling	10:57	11:11	11:33
End of Polling	11:17	12:04	12:11
Ballots/minute, person	2.45	2.38	1.66
Start of Tallying	11:21	12:05	12:17
End of Tallying	11:46	12:21	12:30
Ballots/minute, person	1.96	7.87	4.85
Votes for Winners		75	49
Votes for First Loser		40	76
Observed Risk	.47%	.45%	6.7%

Table 3: Polling Timing. Votes for winners and first loser are not listed for Hartford as the audited race had three winners.

Town	Ballots	Time(m)	# Registrars	Ballots/hour
Hartford	209	14	1	895.7
Norwalk	906	37	1	1,469.18
Hebron	2,723	56	2	1,458.74

Table 4: Manifest Creation Time.

happened in the case of Hartford where the batches revealed one extra ballot while Norwalk and Hebron were spot on (see Table 3). The discrepancy in the case of Hartford was caused by an "ender ballot card" that was left in with the ballots after the election. (An ender card is a specially marked ballot used by the Accuvote system to indicate that tabulation is complete and lock the tabulator.) This was observed during the polling of the ballots. The "ballot" matching the ender card was not selected by the polling but was seen while searching for polled ballots. Such human errors are to be expected and a full process must detail how to handle these situations.

5.2.2 Ballot sample creation

The second stage is driven by the RLA software which, given a risk limit, the margin of victory of the audited race and the manifest, produces the following:

1. the number of ballots to poll,
2. a summary ("ballot pull") sheet showing which ballot from which batch must be polled,
3. for each ballot to be polled, a replacement sheet to be used while polling.

This process prints, for each ballot, a replacement sheet. With a reasonable printer, this takes at most a few minutes to complete and the time is largely negligible.

5.2.3 Ballot sample polling

The third stage is labor intensive for the auditors. It consists of using the summary sheet and processing each pull as follows. At ballot i

1. Use the i^{th} row of the summary sheet to recognize the ballot to pull;
2. Pick up the i^{th} replacement sheet and go to the specified batch;
3. Count from the top of the batch to the specified ballot b ($0 \leq b < 50$) and substitute the replacement sheet for the ballot. Place the pulled ballot at the top of that batch (above the batch cover sheet and any other ballot pulled from that batch before).

When the process ends, several batches have, at their top, the sequence (in order) of the pulled ballot. Table 5

Town	Ballots	Time(m)	# Registrars	Ballots/hour
Hartford	49	20	1	147
Norwalk	126	53	1	142.64
Hebron	126	38	2	99.47

Table 5: Polling Time.

summarizes the timing and show that a weighted average (by size) of

$$\textbf{121 ballots/hour} = \frac{49 + 126 + 126}{20 + 53 + 76}.$$

The pulling process is meant to be reversible (which is to say that the batches can be returned to their original states) so that multiple rounds can be conducted. In the context of the pilot, the plan was for only one round of audits. Registrars took advantage of that knowledge to use a lightweight tabulation scheme that just involved leaving the replacement sheets on top of the batch. (An ideal process would call for returning each ballot to its exact position, though this is not strictly necessary for a second round; certain conventions for polling audits would be satisfied by merely returning batches to the same size.) In addition, we observed that the number of ballots pulled did not match exactly the number of ballots requested. Norwalk pulled 4 fewer ballots than requested with Hebron pulling 2 additional ballots. Such inadvertent errors in sample size can interfere with the risk of the audit: the safest recourse is to simply repeat the selection process to yield a random sample of the right size. To help prevent such sample size mismatches, it may help if auditors carry out online entry of each ballot into audit software rather than pulling all ballots at once. Note that in a statewide audit each municipality would be pulling and tabulating a small number of ballots so this represents minimal overhead.

5.2.4 Ballot sample tabulation

The fourth stage consists of the tabulation of the pulled ballots. In the case of simple races (1 winner only, no cross-endorsement) the process is relatively simple. In the case of more complex races (cross-endorsement, multiple winners) it is slightly more delicate. Both of these settings were observed in the pilot with Hartford's race having three winners[4] and Norwalk's race having cross endorsements.[5] In all cases, it is necessary to determine the number of votes for each candidate. In order to help identify counting errors, it may be useful to tally the number of blanks as well. Table 6 shows the tallying results.

The times here show significant variance. This is a result of the registrars using slightly different methods to tally the results. Both in the case of Norwalk and Hebron, the pulled ballots were collated in a single pile and since the audited race was simple, they were separated into piles (one per candidate and one for blanks) which were then counted. While this is extremely fast, this would not make it possible to return the ballots to their original batches or positions. It seems appropriate to use the slowest estimate from Hartford, i.e., **118** ballots/hour for tallying.

[4]To show the additional complexity, in Hartford, the auditor did hand counting with a tally sheet (while other registrars piled ballots). The registrar calculated the totals three different times, in total, in order to resolve inconsistencies between the first tabulation and the second.

[5]To show the additional complexity, in Norwalk, the auditor tallied using the "pile" method separating in 5 piles: blanks, candidate 1, candidate 2 - endorsement 1, candidate 2 - endorsement 2, candidate 2 - cross-endorsed. He then counted the piles and added up the 3 piles with different endorsements.

Town	Ballots	Time(m)	# Registrars	Ballots/hour
Hartford	49	25	1	117.6
Norwalk	126	16	1	472.49
Hebron	126	13	2	290.76

Table 6: Time to tabulate pulled Ballots.

5.2.5 p-value assessment

The last stage consists of encoding the tallies in Arlo to derive the "p-value" for the experiment to determine the observed risk. A result below the risk limit indicates that the polling process has concluded successfully. The results are reported in Table 3 in the observed risk row. The process takes a negligible amount of time to complete. We note that Hartford and Norwalk had a rather small observed risk. Hebron's observed risk is higher, though still well below the target risk limit. To explain this reference to the "target risk limit": The number of ballots initially chosen by the software to draw in the polling audit is chosen so as to make it likely that the audit will conclude with a p-value below the target value. In the case of Hebron, this target p-value was 10%. Additionally, the extra ballots drawn during sampling may have increased observed risk (see the discussion in Section 5.2.3).

5.2.6 Summary

Some macroscopic observations are relevant for the polling method.

1. The manifest creation was time-consuming. This was done on audit day; hence, there was no need to have a storage solution for the batches. (This would be a further consideration of relevance for a process involving manifest creation on a separate day from the audit.) All the ballots were in a single large container. Operating with small batches makes pulling (and tallying) easy, at the expense of batch creation. It is critical to have the right environment to carry out the manifest creation and holding the batches for the subsequent stages. This feedback was echoed by several registrars. Registrars also remarked on the desire to create a ballot manifest ahead of time at their office which would be done in a full RLA.

2. Storage options should be investigated to be able to build and store ballots according to a manifest for later look up.

3. There were human errors during polling resulting in pulling sometimes more and sometimes fewer ballots than what Arlo requested. This can impact the statistics and should "reset" the process forcing a restart of the polling.

4. The duration of tallying is optimistic. Two of the towns exploited their knowledge of the number of rounds to speed up the tallying process. The slower Hartford numbers are most likely the most realistic for tallying.

5.3 Ballot Comparison Pilot (January 10)

Two towns (Ellington and Stratford) participated in the January 10 comparison-based Risk limited Audit on SotS premises in Hartford. Stratford had two districts and Ellington had one district. Recall that a major goal of the pilot was to test a multi-precinct audit. The registrars brought sealed bags with ballots and the corresponding tapes from the Accuvote terminal.

The audit started at 10 AM with the three precincts in separate (large) rooms. The audits were conducted in five distinct stages. Directions were shared with the registrars at the start of each stage. Stages varied in length once again with the manifest and CVR being the longest. The five stages were:

1. Batch and manifest creation,

2. CVR generation,

3. Ballot sample creation with Arlo (a randomized process),

4. Ballot sample comparison, and

Town	Ellington	Stratford	Stratford
District	2	1	2
Timekeeper	Alex Russell	Laurent Michel	Jack Wohl
Estimated Ballots	867	1561	1019
Requested Risk	10	5	5
% to Complete			
Auditor 1	Wanda DeLand	Louis DeCilio	James Simon
Auditor 2	Lois Timms-Ferrara	James Simon	Abigail Harrison (Voter Center)
Random Seed	46778105257366515170	46407418151903145320	
	32768976198840037439	76000022871017230911	
Actual Ballots	868	1577	1019
Start of Manifest	10:13am	10:13am (90m alone, rest with James)	10:12am (20m 1 person,rest with Abigail)
End of Manifest	11:33am	12:21pm	11:54am
Manifest Questions	+1 ballot w.r.t. tape		
Start audit station	done in parallel with manifest	done in parallel with manifest	done in parallel with manifest
End audit station			
Audited ballots	27	12	7
Start of audit	1:41pm	1:40pm	1:38pm
End of audit	1:59pm	1:58pm	1:51pm
Retrieval Questions			James pulled the ballots, then encoded. In hindsight he preferred Lou's approach (one at a time)
Start of pack up		2:00pm	2:05pm
End of pack up		2:10pm	2:15pm
Discrepancies	None	None	None
Observed Risk	6.3%	2.4%	

Table 7: Comparison Timing

5. p-value assessment.

The remainder of this section details each stage and highlights the key findings. Note that, in general, steps 3-4 need to be repeated (multiple rounds) until the outcome of the round assessment yields sufficient statistical evidence (a $p-$value below the requested risk limit). As for the polling audits, risk limits and sample sizes were chosen to most likely end with a single round.

5.3.1 Batch and Manifest Creation

The process consist of manually batching the ballots in groups of size approximately 50. QR stickers with a single machine and human readable serial number are then placed on the ballots in each batch. The process used sticker application guns commonly used in retail settings. With 5 minutes of practice registrars managed to develop a smooth and efficient process to stick one QR code on each ballot. To simplify the next stages, it is preferable that the QR code on ballots within a batch are sequential (gaps are fine). The QR code can be affixed in any "blank" part of a ballot (front or back).

Difficulties during this stage were caused by stickers infringing on timing marks on the edge of the ballot or QR code stickers crumpling when applied (requiring their removal and application of a new QR code as a ballot can only have one). Out of 3464 ballots, this occurs only a dozen of times in the very first minutes as registrars were learning to use the application guns.

The registrars pipelined this activity with the CVR generation reported next. Hence, the total durations are reported for the composed task only in the next section.

5.3.2 CVR Generation

In this phase, each manually counted and sticker-bearing batch of 50 is submitted to the audit station to accumulate a cast-vote record for the precinct. The audit station reads the QR code and associates it with the record for the ballot. At the end of the process (once all batches have been processed by the audit station), the CVR is exported and uploaded to Arlo for the next phase. Note that the audit station is counting the ballots and interpreting all the marks found on the ballot to produce the CVR of each ballot. This process does not need to be monitored by the registrar and can proceed at the speed of the scanner (and the availability of batches to process).

Since the registrars pipelined the manifest and CVR production, the overall statistics are as shown in Table 8. Note that the duration for the two Stratford precincts were normalized to 1 registrar based on the fraction of time where they worked alone versus when they had assistance from another person.

Town	Ballots	Time(m)	# Registrars	Ballots/hour
Ellington	868	80	2	325.5
Stratford#1	1,577	166	1	570
Stratford#2	1,019	184	1	332.28

Table 8: Manifest, Sticker, and CVR Generation Time.

This summary shows that the combined manifest and CVR creation processed, on average, **408** ballots per hour. This is about 3.5 times slower than the analogous task for the polling method. The extra time is induced by the application of stickers on all ballots.

It is conceivable to consider upgrades to the audit station (and more recent scanners) that side-step part of this time by imprinting a serial number on ballots as they get scanned and using newer scanners that can process at least a ballot per second. Indeed, when considering one precinct, the speed was approximately 6 ballots per minute. While manually counting batches of 50 is identical to polling, the production of the CVR (and imprinting of a serial number) could be done in 30 minutes for 1800 ballots which is faster than the time needed for the manual application of QR codes. We remark on this further in Section 6.

5.3.3 Ballot sample creation with Arlo (a randomized process)

This stage is driven by the Arlo software package which, given a risk limit, the margin of victory of the audited race and the CVR, produces the following:

1. the ballots to compare,
2. a summary sheet showing which ballots from which batches must be pulled, and
3. for each ballot to be pulled, a replacement sheet to be used while pulling.

Once again, this process is fairly fast and is bound by the speed of the printer.

5.3.4 Ballot sample comparison

This stage uses the summary sheet to pull the ballots and enter in Arlo the human interpretation of the marks found on the ballot. For a ballot *i*, the registrar pulls the specified batch and ballot. Retrieving the correct ballot from a batch is made easier thanks to the presence of a serial number (the QR code). Since the verification is done at the ballot level by Arlo, there is no tallying involved at all, eliminating another potential source of human error. The process is reasonably fast as can be seen in Table 9. On average (again, weighted by size), this comes down to **41.2** ballots / hour.

Town	Ballots	Time(m)	# Registrars	Ballots/hour
Ellington	27	18	2	45
Stratford#1	12	18	1	40
Stratford#2	7	13	1	32.31

Table 9: Ballot comparison stage.

Across all precincts, no errors were found when comparing against the relevant CVRs.

5.3.5 p-value assessment

The last stage uses Arlo to derive the "p-value" for the experiment to determine the observed risk. A result below the risk limit is indicative that the ballot comparison process has concluded successfully. The result are reported in Table 7. The process takes a negligible amount of time to complete.

5.3.6 Summary

Note that the time to execute a comparison-based RLA is dominated by CVR creation. This time can be reduced with imprinting, either in primary or transitive tabulators. Further, note that whenever auditing multiple races over the same set of ballots, the race with the tightest margin will approximately determine the number of ballots to be audited; in any case, CVRs only need to created once per election.

Additional observations and questions from registrars Where possible we've included feedback from registrars in discussion of the above procedure. There were some questions about details of the procedure such as requesting ballot manifests to be created at the relevant town hall, providing rubber fingers, and the manner in which instructions are communicated. Here we focus on two pieces of feedback that are relevant for future RLAs.

Manifest Creation Registrars noted the that in both procedures a large fraction of time was spent creating a manifest, as this required hand counting the number of ballots in a batch. This is a step that may be possible to streamline. While an essential feature of ballot manifests is that they are human-verified, the statistical methods discussed in this report can provide strong guarantees even if manifests are not precise, but are just guaranteed to have small errors. This suggests possible methods that may be less time intensive.

Procedure Complexity Multiple registrars noted how much more complicated the RLA procedures are than existing audits.

6 Comparison of Methods

The previous sections describe our research into previous implementations of RLAs and the CT pilot. This section does a more detailed comparison of methods in terms of the difference in setup and execution. This comparison is a snapshot in time and would be greatly affected by new voter-facing tabulators that can imprint and produce CVRs. This section uses state data from the 2020 Presidential Election and a simulation created by the UConn Voter Center. The simulation estimates the number of ballots required for different RLA methods at different margins-of-victory. We use data from the Rhode Island Pilot Program [17] except for the time to sticker ballots and create a CVR using the Audit Station which was not done in Rhode Island. In this section, we focus on a federal year election with a large number of ballots. As we discuss in Section 7, RLAs for local races have different efficiency considerations, and the RLA working group does not recommend any statutory changes for the audit process conducted in municipal years.

- Number of Ballots: 1,823,857 (source:RI)
- Number of Overvotes/Undervotes: 1,824 (source:RI) [1]
- Ballots Counted Per Hour to Create a Manifest: 4,770 (source:RI)
- Ballots Scanned Per Hour to Create a CVR and sticker ballots: 408. See Section 5.3.2 (source:CT)
- Time to Pull a Ballot using Ballot Polling Method: 76 seconds (source:RI)
- Time to Pull a Ballot using Ballot Comparison Method: 45 seconds (source:RI)
- Number of Simulation Runs: 100

6.1 Time to execute a single polling RLA in Connecticut

There are three main pieces of setup necessary to conduct a polling RLA. The first is to organize ballots so that ballots can be easily retrieved. The second is software to conduct the RLA itself. Lastly, create and perform training of registrars on the new methods. Here we focus on ballot manifest and pulling the ballots as these would be done by registrars every year.

Ballot Manifest In addition, no matter the margin a ballot manifest must be prepared (that can be reused across contests). The manifest will take auditors roughly

$$\frac{2.3 \text{ hours}}{\text{municipality}} = \frac{1,823,867 \text{ ballots}}{4,770 \text{ ballots/hour}} * \frac{1}{169 \text{ municipalities}}.$$

Larger municipalities would spend more time in preparation.

	Risk Limit		
Margin	10%	5%	1%
1%	45,283	61,224	91,800
2%	10,939	**16,249**	21,866
3%	5,065	6,562	10,255
4%	2,817	3,770	6,080
5%	1,968	2,537	3,540
10%	520	689	1,161
15%	203	272	422
20%	118	160	219

Table 10: Estimated Number of ballots pulled for Ballot Polling Audit. Ballot manifest can be reused across each yearly audit

Audit Execution As a starting point of comparison, we consider an audit of a single statewide race. Table 10 summarizes the number of ballots pulled for polling methods across different margins and risk limits. We consider a 5% risk limit and 2% margin. Pulling $16,249$ ballots would take

$$\frac{2.0 \text{ hours}}{\text{municipality}} = 16,249 \text{ ballots} * \frac{76 \text{ seconds}}{1 \text{ ballot}} * \frac{1 \text{ hour}}{3600 \text{ seconds}} * \frac{1}{169 \text{ municipalities}}.$$

We stress that the number of ballots pulled in of polling RLAs depends strongly on the margin of the race. A margin of 1.5% would double this number. A strength of ballot polling is the simple setup, little equipment, and relatively simple procedures that are resilient to small errors.

6.2 Time to execute a single comparison RLA in Connecticut

Conducting ballot comparison RLAs is more complicated to enable but becomes more efficient once the necessary setup is in place, especially with smaller margins or multiple audits. Comparison methods depend heavily on the ability to produce a CVR. As mentioned above, this is not a native capability of CT tabulators and was done during the pilot using the VoTeR center audit station. We expect this procedure to become more efficient in the future.

Audit Preparation The pilot used the audit station to generate CVRs with QR stickers for ballot identification (see Section 5.3.2). Each town would need to be equipped with an audit station or tabulation equipment that can produce a CVR and imprint ballots.

Audit Execution The estimated time required for CVR generation and ballot identification using the piloted procedures is

$$\frac{26 \text{ hours}}{\text{municipality}} = \frac{1,823,867 \text{ ballots}}{408 \text{ ballots/hour}} * \frac{1}{169 \text{ municipalities}}.$$

The VoTeR center is working on procedures, methods, and improvements to the audit station to reduce this time. Table 13 shows the number of ballots pulled once the CVR is produced. Pulling 358 ballots would take

$$\frac{.02 \text{ hours}}{\text{municipality}} = 358 \text{ ballots} * \frac{45 \text{ seconds}}{1 \text{ ballot}} * \frac{1 \text{ hour}}{3600 \text{ seconds}} * \frac{1}{169 \text{ municipalities}}.$$

This is roughly 2 minutes per municipality. We emphasize that the time and effort involved in a comparison RLA is largely fixed: in particular it doesn't heavily depend on margin.

6.2.1 New comparison RLAs developed by Voter Center

In the above, we are confident in the estimate of the time for polling RLAs and less confident in time to produce a CVR for the comparison RLAs. The VoTeR center is actively working on improvements to this step of the process. In addition, the VoTeR center has created new auditing methods that allow for only a fraction of the CVR to be produced.

	Risk Limit		
Margin	10%	5%	1%
1%	582	763	1,178
2%	272	**358**	536
3%	178	232	352
4%	130	173	261
5%	103	137	211
10%	52	68	104
15%	34	45	69
20%	26	34	51

Table 11: Number of ballots sampled for ballot comparison audit.

We call these methods lazy ballot comparison audits. For a single audit, at a 5% and 2% margin these methods require less than half of the CVR to be produced.

6.3 Time to execute multiple RLAs in Connecticut

Polling methods offer good efficiency unless a race occurs with a small margin, with comparison methods requiring up-front investment but adapting well to all auditable margins and auditing multiple races. In particular, suppose we conduct 15 audits each election with different populations (continuing to adopt a 2% margin and 5% risk limit); then the overall time to conduct all audits is:

Polling: 32.3 hours = 2.3 hours to setup + $15 * 2$ hours to conduct the RLA

Comparison: 26.3 hours = 26 hours to setup + $15 * .02$ hours to conduct the RLA

Remarks. Note that the ballot manifest and any generated CVRs can be reused for multiple audits. Auditing multiple races with the same ballot population can be done with significant improvements in efficiency. However, if one wishes to audit races over different populations (as one would expect in typical election years) one must pull sufficient ballots from each population.

7 Conclusion and Recommendation

Implementing an RLA in CT would require new procedures, equipment, and time commitment on the part of CT registrars. However, these methods provide high confidence that the reported outcome of elections agrees with voter intent. Thus, the RLA working group recommends that the State of Connecticut implement RLAs. Specifically, the RLA working group makes the following recommendation:

> The State of Connecticut should adopt comparison risk-limiting election audits for all federal year elections. These risk-limiting audits should be managed and administrated by the Office of the Secretary of the State. All statewide races should be audited; in addition, at least one US House of Representatives race should be audited and at least 5% of state legislative races. Risk limits for each audit should be no more than 5%. Ballot manifests should be created for all ballots cast in all state elections including municipal year elections.
>
> The Office of the Secretary of State should be responsible for annual reports describing the audit procedures and results. The State should continue 1) its technological audits of removable media used to program and collect results for Connecticut tabulators and 2) current auditing procedures for municipal year elections.

Rationale The above recommendation primarily consists of three parts (i.) Creation of ballot manifests for all elections, (ii.) Adopting RLAs in all federal year elections in CT, and (iii.) Use of comparison RLAs. We provide justification for these recommendations separately.

1. **Ballot manifests.** All current RLA methods rely on the creation of ballot manifests. We remark that the most significant source of discrepancies in previous equipment audits has been mismatches between the number of audited ballots and the number of tabulated ballots. Creating a ballot manifest immediately after an election may make it easier to reconcile these discrepancies. Even if Connecticut does not decide to adopt an RLA, insisting on ballot manifests is a fairly lightweight auditing convention that can identify problems. As ballot manifests require careful processing of ballots after an election, they can also lead to more dependable ballot handling and storage.

2. **Adopting RLAs.** Risk-limiting audits provide a rigorous, conspicuous mechanism to establish trust in election outcomes. Furthermore, they can provide these guarantees without a full recount and can offer transparent implementations that support verification by external observers.

3. **Comparison Audits.** The efficiency of RLAs depends on procedures, statistical analyses, margins, selected races, and equipment/software. Comparison RLAs provide the best efficiency for small margins and scale well to auditing of multiple races. As small margins are those situations where it is most important to verify results, efficiency for such races is a priority. Such methods also have an opportunity for improvement in efficiency as equipment evolves, while polling methods are unlikely to improve substantively.

Required changes in statute. The RLA working group expects there to be several statutory changes necessary to support these changes. In particular, the current audit procedure is encoded in CT statutes. In addition to procedural changes and providing the Office of the Secretary of State with the proper authority, RLAs have two other important changes: transparency and force of law.

Transparency For RLAs to be transparent and open, they make publicly accessible several pieces of information that are not currently available to the public. This includes tabulator subtotals, CVRs, and photos of sampled ballots.

Force of Law RLAs must be explicitly called for by law, must be conducted before the election certification, and must be accompanied by a clear procedure of how to proceed if an RLA is unable to gather sufficient evidence that the election is valid. We recommend that such an outcome shall lead to a full recount of the contest in question.

References

[1] Mark Lindeman and Philip B Stark. A gentle introduction to risk-limiting audits. IEEE Security & Privacy, 10(5):42–49, 2012.

[2] Jennie Bretschneider, Sean Flaherty, Susannah Goodman, Mark Halvorson, Roger Johnston, Mark Lindeman, Ronald L. Rivest, Pam Smith, and Phillip B. Stark. Risk-limiting post-election audits: Why and how, 2012.

[3] Kellie Ottoboni, Philip B Stark, Mark Lindeman, and Neal McBurnett. Risk-limiting audits by stratified union-intersection tests of elections (SUITE). In International Joint Conference on Electronic Voting, pages 174–188. Springer, 2018.

[4] Philip B Stark. Sets of half-average nulls generate risk-limiting audits: Shangrla. In International Conference on Financial Cryptography and Data Security, pages 319–336. Springer, 2020.

[5] Jennifer Morrell. Knowing it's right, part two. risk-limiting audit implementation workbook., 2019.

[6] Matthew Bernhard. Risk-limiting audits: A practical systematization of knowledge. In International Joint Conference on Electronic Voting, 2021.

[7] Philip B Stark. Auditing a collection of races simultaneously. arXiv preprint arXiv:0905.1422, 2009.

[8] Philip B Stark. Super-simple simultaneous single-ballot risk-limiting audits. In EVT/WOTE, 2010.

[9] Michael J Higgins, Ronald L Rivest, and Philip B Stark. Sharper p–values for stratified election audits. Statistics, Politics, and Policy, 2(1), 2011.

[10] Mark Lindeman, Philip B Stark, and Vincent S Yates. Bravo: Ballot-polling risk-limiting audits to verify outcomes. In EVT/WOTE, 2012.

[11] Mark Lindeman and Neal McBurnett. Risk-limiting audits by stratified union-intersection tests of elections (suite). E-Vote-ID 2018, page 294, 2018.

[12] Kellie Ottoboni, Matthew Bernhard, J Alex Halderman, Ronald L Rivest, and Philip B Stark. Bernoulli ballot polling: a manifest improvement for risk-limiting audits. In International Conference on Financial Cryptography and Data Security, pages 226–241, 2019.

[13] Ian Waudby-Smith, Philip B Stark, and Aaditya Ramdas. Rilacs: Risk limiting audits via confidence sequences. In International Joint Conference on Electronic Voting, pages 124–139. Springer, 2021.

[14] Michelle Blom, Jurlind Budurushi, Ronald L Rivest, Philip B Stark, Peter J Stuckey, Vanessa Teague, and Damjan Vukcevic. Assertion-based approaches to auditing complex elections, with application to party-list proportional elections. In International Joint Conference on Electronic Voting, pages 47–62. Springer, 2021.

[15] Tigran Antonyan, Theodore Bromley, Laurent Michel, Alexander Russell, Alexander Shvartsman, and Suzanne Stark. Computer assisted post election audits. State Certification Testing of Voting Systems National Conference, 2013.

[16] Colorado Secretary of State. 2020 general election risk-limiting audit discrepancy report, 2020.

[17] Brennan Center for Justice and Rhode Island RLA Working Group. Pilot implementation study of risk-limiting audit methods in the state of Rhode Island, 2019.

[18] Risk limiting audits with Arlo, 2021.

[19] Jennifer Morrell and Democracy Fund Election Validation Project. Putting election integrity to the test: A case study of new jersey's first pilot of risk-limiting audits, 2019.

[20] Secretary of the State of Connecticut. Audit procedure manual, 2021.

[21] Election Assistance Commission. Recounts and audits study.

[22] Jerome Lovato. Risk-limiting audits - practical application, 2018.

[23] California Secretary of State. Post-election risk-limiting audit pilot program 2011-2013: Final report to the united states election assistance commission, 2013.

[24] El Dorado County. Certification of contest results and report of audit results, 2020.

[25] Inyo County. Certification of contest results and report of audit results, 2020.

[26] Mark Lindeman. Rhode Island presidential risk-limiting audit, november 19-24, 2020 (brief report), 2020.

A Procedure Overview

You've volunteered to help the state of Connecticut pilot a risk-limiting audit (or RLA, for short). These are procedures that provides mathematical guarantees on the chance that the reported election results do not match the cast paper ballots. Your experience here today will be used to inform a report delivered to the Connecticut legislature. The goal is to understand the unique challenges and opportunities for Connecticut and ultimately make a recommendation on the value proposition of RLAs.

We will be piloting the two most commonly used RLAs, ballot polling on January 6 and ballot comparison on January 7. The software we will be using for both methods is Arlo `https://www.voting.works/risk-limiting-audits` which is produced by Voting Works `https://voting.works/`. Discussion in this document is split up into procedures for the two types of audits with Section C discussing ballot polling and Section D discussing ballot comparison. However, both audits share a first step of creating a ballot manifest: a hand counted and batched collection of the paper ballots. This is discussed in Section B. For this pilot we will be auditing a single race.

Before beginning make sure that you have:

1. A set of ballots corresponding to the selected precinct(s),

2. A tabulator tape with subtotals corresponding to the ballots.

All other needed documentation will be produced as part of the audit. VoTeR center staff and the Secretary of State (SotS) office will be serving as the State level administrator for this audit including launching the audit. Your role will be to prepare ballots, locate ballots when they are selected for audit, and interpret the results. Depending on the type of audit these will either be recorded manually on paper or into audit software. VoTeR center staff will be observing and collecting statistics. Please communicate any part of the process that you find easy, hard, or noteworthy.

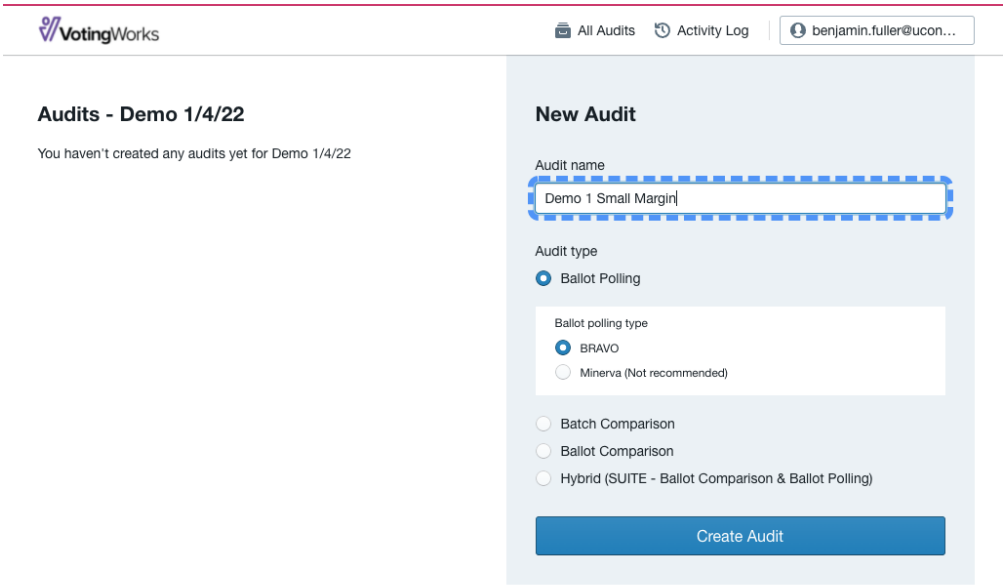

Figure 1: Audit Creation

Live Demonstrations

These demonstrations will be conducted by VoTeR center staff from 10-11:30am on your audit day.

For ballot comparison (January 7):

- Manifest creation, open Excel on laptops. Show how to save something as CSV. Prepare sample files.
- Sticker application with sticker gun
- CVR creation

Arlo Screenshots

Jurisdiction Manager: Logging into Arlo. If we have to create an admin account:

1. Use URL: http://voter-automation.voterlab:3000/auth/support/start with appropriate hostname and port.
2. Once here enter an arbitrary email.
3. Create an audit organization and audit administrator with our email.

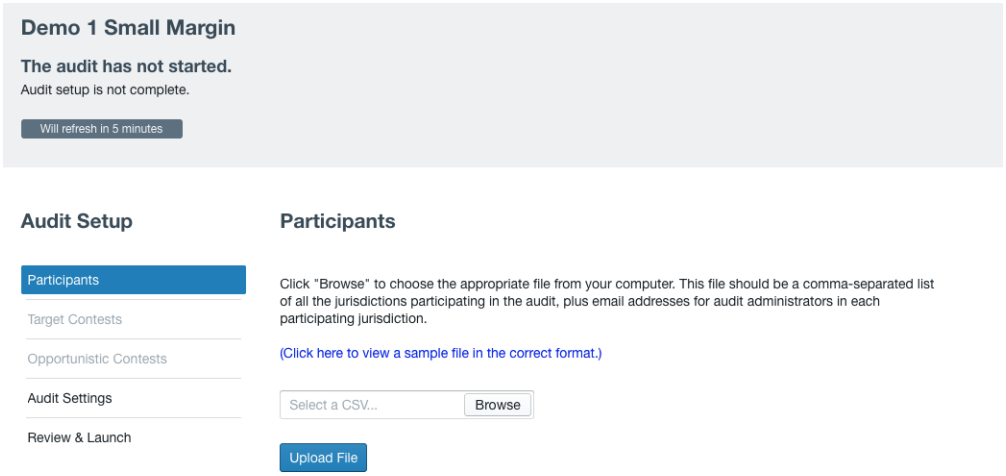

Figure 2: Upload Participants Files (should be done before audit starts).

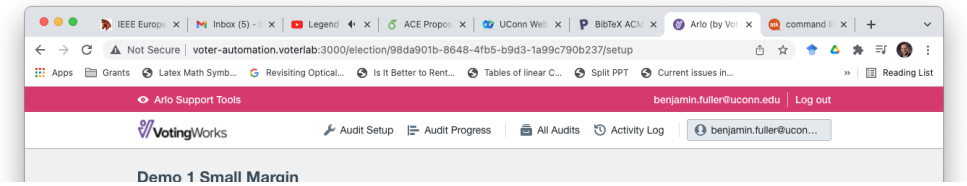

Figure 3: Switch Role.

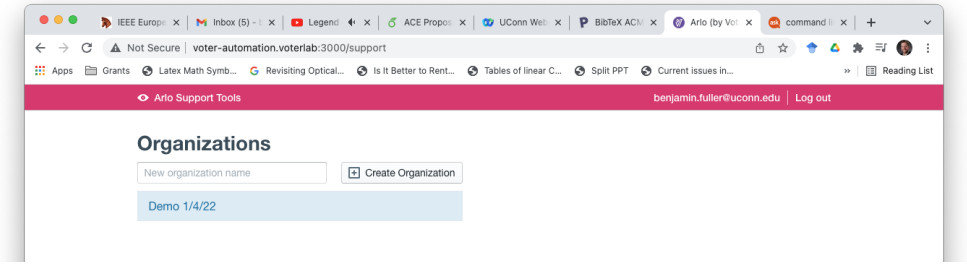

Figure 4: Switch Role Step 2.

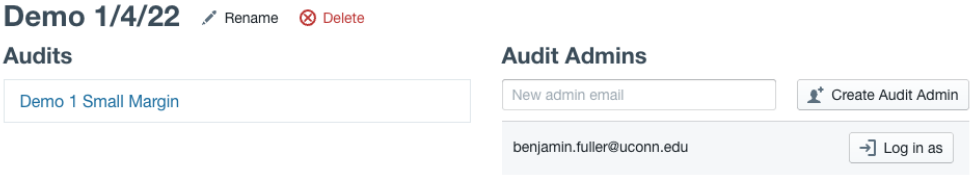

Figure 5: Switch Role Step 3, from here you can act as any individual role in the audit.

B Ballot Manifest Procedure

Important Note: Only use one ballot type for the audit; if you have more than one ballot type, please consult someone from the Voter Center.

1. Collect all of the ballot boxes for your precinct in one room.
2. For each ballot box, hand count the number of ballots in that box.
    - If conducting a ballot polling audit, then after counting 50 ballots, set those ballots aside in a separate container to create a batch. After creating a batch, add cover sheets to the batches sequentially (i.e. Batch 1, Batch 2, etc.).
    - If conducting a ballot comparison audit, then after counting 50 ballots, set those ballots aside in a separate container to create a batch. After creating a batch, label the batches sequentially (i.e. Batch 1, Batch 2, etc.). As you hand count, you will be applying stickers to the back of each ballot. A demonstration on how to do this will be shown before the audit starts.
3. Record the number of ballots per batch in the ballot manifest template provided. Create new rows per batch as necessary. When you are finished, save this as a `.csv` file. VoTeR center staff will show you how to export the file. On ballot polling day this will be manually communicated to VoTeR center staff.
4. Add the total number of ballots counted from your ballot manifest. Note any discrepancies between your hand counted total and the tabulator tape and inform VoTeR center staff.

C Ballot Polling Procedure

There are four major stages to the ballot polling RLA. These are:

1. **Ballot manifest creation**. **Auditor** A hand counting of the number of ballots. See Section B for information on creating the ballot manifest.

2. **Ballot manifest upload**. **Voter center** Creation of an Excel spreadsheet with the ballot manifest which is loaded into Arlo.

3. **Audit launch**. **Voter center** Randomly select ballots to examine. We want 60 bits of security: That's 14 rolls of a 20-sided die or 24 rolls of a six-sided die.

4. **Ballot retrieval and recording**. **Auditor** Pulling the selected ballots and keep a hand total of the votes on the selected ballots. After all ballots are retrieved and recorded, these will be hand communicated to Voter center staff which will enter them into audit software.

C.1 Ballot retrieval and recording

1. **Voter center** enter ballot manifest into Arlo, tabulation should already be done. After that should be good to initialize audit.

2. **Voter center** print out selected ballots and placeholder sheets. Walk these to the room with registrars.

3. **Auditor** you will receive a ballot retrieval sheet with the batch name and ballot number for each ballot you are expected to retrieve. In addition, you will receive placeholder sheets for each retrieved ballots. Look at the first batch in the ballot pull sheet, note all of the ballots that need to be pulled from each batch. Perform the following:

   (a) Retrieve the corresponding ballot from the batch counting from the top of stack,

   (b) Once all ballots from the stake are retrieved place the placeholder sheets on the top of the stack, one for each ballot retrieved.

   Repeat this process for each batch you need to retrieve ballots from. At the end of this process, you should have used all placeholder sheets.

   Once you have retrieved all of the relevant ballots, you can begin the interpretation. Record the vote subtotals per candidate on the ballot pull sheet. You can do this either by sorting ballots into piles for each candidate for a simple election or using the tally sheet.

4. **Auditor** Verify you have entered the correct results before submitting the audited ballots.

   Once you have interpreted all the ballots, provide VoTeR center staff with the tally sheet.

5. **Auditor** Replace all selected ballots into batches with each selected ballot replacing one placeholder sheet. You do not need to place these ballots into the batches they started from.

6. **VoTeR center** Enter subtotals into Arlo and calculate risk. May need to launch another round.

7. **Auditor** After step 6, await further instruction.

D Ballot Comparison Procedure

There are six main stages to a ballot comparison RLA. These are:

1. **Ballot manifest creation**. **Auditor** A hand counting of the number of ballots. See Section B for information on creating the ballot manifest.
2. **Ballot manifest upload**. **Auditor** Creation of an Excel spreadsheet with the ballot manifest which is loaded into Arlo.
3. **Cast Vote Record Creation**. **Auditor** Creation of a full listing of all votes in the batch. This will be created using the VoTeR Audit Station.
4. **Cast Vote Record Upload Auditor**
5. **Audit launch**. **Voter center** Results in some set of ballots to examine.
6. **Ballot retrieval and entering**. **Auditor** Pulling the selected ballots and entering the votes on each ballot into Arlo. If necessary and time allows a second round of audit may be needed.

D.1 Ballot manifest upload (Auditor)

1. Follow the Arlo Documentation for how to log in to your jurisdiction administrator account and upload the necessary pre-audit files.
2. After the necessary files are uploaded, the audit administrator can begin the audit. Select the number of audit boards participating in this audit. The number of ballots to retrieve will be distributed between the number of audit boards.
3. Once the audit boards are created, each audit board will receive a QR code and link to log in to their account. Using the link will automatically take you to your audit board page.

D.2 Cast Vote Record Creation (Auditor)

1. Before we can begin a ballot comparison audit, we need to create a CVR using the Voter Center audit station. We will demonstrate how to do this before the audit starts. Once you have created a CVR, save the file as "CVR_YourTownPrecinctNumber.csv" (such as "CVR_Hartford1.csv").
2. Ask the VoTeR center staff to convert the CVR to a form that can be used by Arlo.

D.3 Cast Vote Record Upload (Auditor)

1. Follow the Arlo documentation for how to log in to the jurisdiction administrator account and upload the necessary pre-audit files. For a ballot comparison audit, this would be a ballot manifest and CVR.
2. After the necessary files are uploaded, the audit administrator can begin the audit. Select the number of audit boards participating in this audit. The number of ballots to retrieve will be distributed between the number of audit boards.

D.4 Ballot retrieval and entering (Auditor)

1. Once the audit boards are created, each audit board will receive a QR code and link to log in to their account. Using the link will automatically take you to your audit board page.
2. Type in your name on the Audit Board: Member Sign-in page and then click Next to continue.
3. Your audit board page will have a ballot retrieval list and the respective ballot placeholders. Print out this ballot retrieval sheet as well as the placeholders.

4. For each ballot on the retrieval list, go to the corresponding batch. Count through the ballot number starting from either the front or the back, whichever would be quicker. Pull out the ballot and compare the QR code to the Imprinted ID column on the ballot retrieval sheet. The imprinted ID has the following format: Town-BatchNumber-QRNumber, so if the last number of the imprinted ID matches the QR code number, then you pulled the correct ballot. If this is the case, replace the ballot location in the batch with the corresponding ballot placeholder sheet. If the QR code number and imprinted ID number do not match, then replace the ballot you currently have and count either up or down until you reach the correct ballot. Repeat this process until you have retrieved all the ballots you need from this batch.

5. Interpret the specific contest results of each ballot and record them next to the corresponding ballot in the ballot retrieval list.

6. Repeat steps 4-5 for each batch in the retrieval list.

7. Once you have interpreted all your ballots, enter each ballot's results into Arlo. Verify you have entered the correct results before submitting the audited ballots.

8. After you submit, each audit board member should type in their name as their signature on the audit. You must type in your name the same way you did when you created your sign-in.

9. After step 9, await further instruction.

Town	Hartford	Norwalk	Hebron
District	7	E-1	1
Timekeeper	Alex	Laurent	Ben
Estimated Ballots	208	906	3000
Requested Risk			
% to Complete			
Auditor 1			
Auditor 2			
Random Seed			
Actual Ballots			
Start of Manifest			
End of Manifest			
Manifest Questions			
Audited ballots			
Start of Polling			
End of Polling			
Polling Questions			
Start of Tallying			
End of Tallying			
Start of Packup			
End of Packup			
Votes for Winners			
Votes for First Loser			
Observed Risk			

Table 12: Polling Timing

Town	Ellington	Stratford	Stratford
District	2	1	2
Timekeeper	Alex	Laurent	Ben
Estimated Ballots	868	1561	1019
Requested Risk			
% to Complete			
Auditor 1			
Auditor 2			
Random Seed			
Actual Ballots			
Start of Manifest			
End of Manifest			
Manifest Questions			
Start audit station			
End audit station			
Start launch audit			
End launch audit			
Audited ballots			
Start of retrieval			
End of retrieval			
Retrieval Questions			
Start of Data Entry			
End of Data Entry			
Retrieval Questions			
Start of Packup			
End of Packup			
Discrepancies			
Observed Risk			

Table 13: Comparison Timing

E Case Studies from Other States

Colorado: 2017 Coordinated Election [22]

- RLA methods: Ballot-Level Comparison and Ballot Polling
- Number of Counties Participated: 56
- Biggest Successes: Open communication with county clerks resulted in the best practices being developed, replicating ballot content on RLA software simplified the comparison audit, tally spreadsheets helped those with ballot-polling audits, RLA was completely transparent.
- Biggest Challenges: Communication of RLA information from election officials to the public and stakeholders, maintaining ballot order of scanned ballots, provide sufficient time for training of election officials and testing of software.
- Recommendation: Colorado will conduct RLAs for the 2018 Primary and General Election. Now required by law.

California: 2011-2013 Pilot Program [23]

- RLA methods: Ballot-Level Comparison, (Small) Batch-Level Comparison, Ballot Polling, Transitive
- Number of Counties Participated: 14
- Biggest Successes: RLA was more effective than manual counting ballots from 1 percent of precincts, step-by-step audit instructions were developed for future use.
- Biggest Challenges: Voting systems did not export data in software-friendly format, voting systems did not report results in batches (reported by precinct), chain-of-custody issues with the ballots.
- Recommendation: Allow counties to conduct RLAs in lieu of the 1 percent manual tally and conduct ballot polling audits for large contests.

California: 2020 Presidential Election [24, 25]

- Number of Counties Participated: 2 (El Dorado and Inyo County)
- Biggest Successes: Both audits conducted publicly with full transparency.
- Biggest Challenges: Data incorrectly entered into software; both counties made vendor recommendations to improve ease of use.

Rhode Island: 2019 Pilot Program [17]

- RLA methods: Ballot-Level Comparison, Ballot Polling, Batch-Level Comparison
- Number of Jurisdictions Participated: 3 (Bristol, Portsmouth, Cranston)
- Biggest Successes: Pilot methods worked well to investigate discrepancies, each method worked successfully with as few ballots as possible, pilot was open to general public/transparent.
- Biggest Challenges: Retrieving the correct ballot, lack of proper training and management of auditing staff, location for ballot storage and containers.
- Recommendation: Implement a ballot-level comparison RLA.

Rhode Island: 2020 Presidential Election [26]

- RLA methods: Batch-Comparison and Ballot Polling
- Rhode Island opted for these methods (which differ from the pilot recommended) as a CVR is not necessary.

New Jersey: 2018 Pilot Program [19]

- RLA methods: Ballot-Level Comparison, Ballot Polling, and Transitive
- Number of Counties Participated: 5
- Biggest Successes: Greater transparency with RLA software, RLA was conducted quickly and efficiently, the development of a ballot manifest ensured ballot accountability.
- Biggest Challenges: Disagreement over voter intent, issues with paper-ballot handling and storage (most data on memory cards).
- Recommendation: Develop a collaborative working group to recommend election rules and statutory changes to advance risk-limiting audits.