Comprehensive Risk Estimation in Election Audits

And why Ballot Image Audits should be our primary goal

Ray Lutz, Citizens Oversight. 2019-08-15 raylutz@citizensoversight.org

This paper can be found at: http://www.copswiki.org/Common/M1913

This paper considers a comprehensive risk calculation for election systems that use hand-marked paper ballots.

Most of the "Risk Limiting Audit" calculations for estimating the risk presented in frequently cited literature (such as "Super-Simple Simultaneous Single-Ballot Risk-Limiting Audits," by Philip B. Stark[1], also described in "A Gentle Introduction to Risk-limiting Audits," by Mark Lindeman and Philip B. Stark[2], "Bayesian Tabulation Audits Explained and Extended," by Ron Rivest[3], and for polling audits: "BRAVO: Ballot-polling Risk-limiting Audits to Verify Outcomes", by Lindeman, Stark and Yates[4], and "Clip Audit," by Ron Rivest[5].) focus on a very specific and limited scope while relying on low or zero risk for other aspects. These methods sometimes suggest that they will limit the risk that the election outcome is incorrect to some specific value, sometimes shrouded in law. Unfortunately, the calculations presented do not provide a comprehensive calculation and underestimate the actual risk. In some cases (particularly with poor implementation), such audits may not limit the risk at all.

Before continuing, a quick review of the various types of tabulation audits is in order.

Ballot Polling Audit – Pulls random ballot samples and compares the margin with the official margin. This is similar to polling done to estimate how people might vote, and it relies on well understood statistics. This results in a large number of ballot samples, particularly if the margins are close (say < 10%), and sampling individual ballots is onerous and requires steady oversight.

Ballot Comparison Audit – Pulls random ballot samples and compares each with the cast vote record for that ballot. This is the most onerous auditing process, requires ballot-level granularity in the cast-vote record, and become unwieldy with small elections and/or close margins (say < 1%), and are difficult to understand and oversee. But they are the most efficient of the sampling RLAs in terms of the number of ballots inspected at any given margin and risk limit.

Batch Comparison Audit – Pulls random batches and hand tallies them, and then compares with the official result for that batch. This is very suitable in large elections with simple ballots and with

[1] https://www.usenix.org/legacy/events/evtwote10/tech/full_papers/Stark.pdf
[2] https://www.stat.berkeley.edu/~stark/Preprints/gentle12.pdf
[3] http://people.csail.mit.edu/rivest/pubs/Riv18a.pdf
[4] https://www.usenix.org/system/files/conference/evtwote12/evtwote12-final27.pdf
[5] http://people.csail.mit.edu/rivest/pubs/Riv17b.pdf

margins > 3% and relatively low risk limits (say 20%[6]), but again not good for small or close elections. Keeps the ballots in batches, which is nice.

Ballot Image Audit – Images are created of each ballot and secured by publishing cryptographic signatures as soon as possible after scanning and preferably before they are evaluated[7]. The images should be made public so that third parties can perform their own evaluation of the images. Images should be validated by comparing with the corresponding paper ballots. This is best performed by a separate team, creating a Quality Control report for randomly chosen ballots in each batch imaged. The images are then processed by software or manual inspection to generate the audit tabulation, which can be compared with the official tabulation.

This is the most open of the options, which is very satisfying to voter confidence, and supports "crowd audits" where the public reviews the images to the ballots in those close races. Image validation is recommended, but if not performed, the performance of this type of audit with regard to overall confidence is still competitive. Ballot Image Audits have been used in Maryland and evaluated in comparison with other methods. Clear Ballot is a private firm that offers third-party ballot image audits and other open-source options will likely become available.

To get an idea of the comprehensive risk, a risk or "confidence factor" is estimated for each step of the process. The confidence factor is directly related to the risk. The risk is (1 – confidence factor).

The total confidence is the product of all the confidence factors. The following table describes each of the confidence factors Cn. These confidences are with respect to an outside observer.

Cn	Risk Step	Ballot Polling Audit	Ballot Comparison Audit	Batch Comparison Audit	Ballot Image Audit (BIA)
C1	Confidence that ballots are NOT modified / added / deleted prior to scanning	All processes rely on a robust chain of custody prior to scanning the ballots. Audit procedures should review the number of ballots cast at polling places and by mail to ensure all ballots are included and invalidated ballots are minimized.			The same nonzero risk unless voter reviews and verifies the ballot image.

[6] The risk limit calculation is generally very conservative in batch comparison audits, because it is difficult to predict exactly how an aversary might affect the batches without causing immediate suspicion. Thus, 20% might be equivalent to 10% risk in other approaches if the predicted hack is is allows which may not really be possible given reported results.

[7] A standard approach for ballot image security will need to be established, but the process is essentially publishing appropriately signed cryptographic hashes of the images. These techniques are well understood and easy to implement, and will make it infeasible to change ballot images without the possibility of detection.

<table>
<tr><th>Cn</th><th>Risk Step</th><th>Ballot Polling Audit</th><th>Ballot Comparison Audit</th><th>Batch Comparison Audit</th><th>Ballot Image Audit (BIA)</th></tr>
<tr><td>C2</td><td>Confidence that images are NOT modified / added / deleted after scanning ballots and prior to securing images.</td><td colspan="3">0% risk. However, images may be used to reduce risk that ballots are modified prior to sampling or used to conduct any requirement for a “full-hand count” and then the risk element would be included for that aspect.</td><td>Nonzero risk[8] can be mitigated by comparing images to paper, and reduced by minimizing window of opportunity</td></tr>
<tr><td>C3</td><td>Confidence that ballots are NOT modified after scanning and prior to audit sampling.</td><td colspan="3">Nonzero risk – Ballots can be modified after scanning to cover up hacked tabulation</td><td>0% risk - does not rely on paper ballots or sampling</td></tr>
<tr><td>C4</td><td>Requires frozen (and fully published) Cast-Vote-Records</td><td>0% risk – does not require a cast vote record but does rely on reported margin</td><td colspan="2">Nonzero risk – If CVR is not properly frozen means audit can be defeated by unmodifying hacked entries that are selected for audit so they will match the ballots.</td><td>Does not need a CVR but may want to compare with the official CVR. 0% risk</td></tr>
<tr><td>C5</td><td>Random Selection</td><td colspan="3">Requires oversight of random number procedure</td><td>0% risk except when used for image validation</td></tr>
<tr><td>C6</td><td>Drawing samples</td><td>May have 100% risk, audit can be defeated by choosing ballots that support desired ballot option</td><td>May have 100% risk, audit can be defeated by choosing ballots that match the CVR</td><td>Pulling of audited batches is simple as they can remain in sealed boxes until audited but audit can be defeated if batches are chosen that match CVR</td><td>0% risk except when used for image validation</td></tr>
<tr><td>C7</td><td>Reliant on a ballot manifest</td><td>Usually, and if so, non-zero risk</td><td>Yes - non-zero risk as manifest may not include all ballots</td><td>No, 0% risk - batches are audited as stored</td><td>No, 0% risk</td></tr>
</table>

[8] An example of a hack that could modify ballot images was investigated and presented in the paper “UnclearBallot: Automated Ballot Image Manipulation” http://kartikeyakandula.com/unclearballot.pdf

Cn	Risk Step	Ballot Polling Audit	Ballot Comparison Audit	Batch Comparison Audit	Ballot Image Audit (BIA)
C8	Data entry of ballots may be manipulated by custom software or entry may be subject to “innocent fix-up”	non-zero if custom DRE-like[9] audit software is used. Can be mitigated by using standardized tally sheets that are completed without computer assist so data entry can be validated. “Innocent fix-up” risk can be reduced with careful procedures.			0%, as no data entry is required as images are used directly
C9	Compatible with sealed ballots due to possible judicial contest and court order to that effect.	Non-zero risk - Audit may not be able to be performed if such a court order exists.		Somewhat - batches need not be compromised	Yes -- Image audit can be performed without obtaining access to ballots and may reduce judicial contests
C10	Compatible with third party and competitive auditing	No	No	No	Yes – The public can conduct their own audits or count the ballot images by hand.
C11	Relies on expensive full hand count if audit finds problems	Yes, plus if full hand count is used, testing indicates a 1% to 2% error rate			No. 100% Audit is deterministic and is like a full hand count for all races.
C12	May confirm a hacked election	Yes. This is the normal risk cited in Risk-Limiting Audit procedures. For batch-comparison audits the risk is more difficult to estimate to provide optimal sampling.			0%. Deterministic procedure does not include risk at this stage.
C13	May not include all contests at the same risk limit	Yes. Implementation of RLAs typically is for state-wide and county-wide contests only; other contests do not receive risk-limited treatment.			0%. All contests are treated equally with 100% review
C14	Confidence that the result of the audit is accurately reported	Any risk can be nearly fully mitigated if a standard audit report format is utilized, and it is possible to verify the content of the report by analysis of other data provided.			0% as the report is not relied upon if ballots are published.

[9] DRE is the term used to describe “direct recording electronic” voting machines where the use makes their selections and it is stored in memory, and there is no audit trail to check whether the values were correctly entered into memory. This sort of machine is now known to be problematic for voting due to the lack of the audit trail. If the software used for auditing acts like a DRE in that a paper trail, such as a tally sheet, is not also used, then we cannot be certain that the software entered the data correctly. For this reason, the use of a tally sheet as the primary or back up input mechanism is required to avoid this risk.

We obtain an equation of probability similar in form to the famous "Drake Equation" which is used to estimate the likelihood that we might interact with extraterrestrial life. Each of the confidence factors above will give a value 0 to 100% confidence.

Thus, the overall confidence is

C(comprehensive) = C1*C2*C3*C4*C5*C6*C7*C8*C9*C10*C11*C12*C13*C14

Note that if any one step provides low or zero confidence, this reduces the overall confidence to that level. Granted, the numerical values of these confidences will be estimates, but we can include or exclude certain risk factors fairly easily.

The sampling-based RLA types require most of the confidence factors, as follows. (C2 is never included because ballot images are not used.)

Ballot Polling Audit:

C(ballot polling) = C1*C3*C5*C6*C7*C8*C9*C10*C11*C12*C13*C14

Ballot Comparison Audit:

C(ballot comparison) = C1*C3*C4*C5*C6*C7*C8*C9*C10*C11*C12*C13*C14

Batch Comparison Audit:

C(batch Comparison) = C1*C3*C4*C5*C8*C10*C11*C12*C13*C14

Before continuing, it is also fair to say that the papers cited above (in footnotes 1 through 5) do mention some of these risks and do attempt to mitigate and minimize some of them through procedures. Unfortunately, it is also the case that in many of the pilots and existing implementation of these audits, some of the risk factors are excessive due to poor procedures. As an example, in Colorado, not all races are chosen and those that are chosen are not chosen randomly (C5 = 0%, C13 maybe 0% for some contests), and they do not make the Cast Vote Record public (C4 = 0%) so it is not possible to provide sufficient oversight. In the pilots in Orange County, CA, they did not provide any transparency to the ballot sampling process (C6=0%) in their ballot polling audit. In Los Angeles, they rerun any batch and the only compare with the new report, and do not report the true discrepancies (C14=0%).

For **Ballot Image Audits**, (for now without image validation) many of these confidences are 100% (0% risk), assuming the procedure includes securing the ballot images properly, and publishes the ballot images.

C3 (Uses the images prior to this stage that are already secured)

C4 (Does not use a CVR, although some methods do compare with it.)

C5 (Does not rely on random number selection)

C6 (Does not draw samples)

C7 (Does not use ballot manifest)

C8 (Does not use data entry in audit process)

C9 (Does not conflict with court orders)

C10 (Audit team can be checked with redundant competitive audits)

C11 (Does not use manual tally procedures)

C12 (Does not have a risk in not detecting a hacked election as 100% of ballots are inspected)

C13 (All contests are included in the audit and reviewed 100%)

C14 (All ballots are published, so there is no reliance on the audit report.)

So, for Ballot Image Audits:

C(comprehensive, BIA) = C1*C2

We are left with C1, confidence that ballots are NOT modified / added / deleted prior to scanning, which all methods must endure[10], and C2, the confidence that images are NOT modified / added / deleted after scanning ballots and prior to securing images.

For many simple hacks, such as modifying only the Cast Vote Record, they will be caught regardless of whether the images are modified after being scanned but before being tabulated, because the hack is done only to the cast vote record. To avoid detection if ballot images are available, it would require changing the images prior to being secured (C2), which is a much more difficult hack.

Before we move on to deal with reducing C2 by performing image validation, there are some other more specific risks that are unique to BIAs and would not be detected by a single pass of automated parsing (i.e. extracting the vote) from the images. These issues would, however, be detected with further inspection and competitive audits, which the other types don't offer, i.e. C10.

For example, let’s say that the x,y coordinates of the bubble to be darkened were swapped with the other ballot option. That would effectively swap the votes for the two candidates. Automated ballot parsing should not rely solely on style information (x,y coordinates of the locations of ballot options) from the election system. The human-readable text should be processed by the image parsing software or human inspectors, rather than just looking at the bubble itself based on x,y values. That would mitigate this risk somewhat without competitive audits.

Parsing of hand-marked paper ballots will include differences with regard to how voter intent should be interpreted. BIA software can also detect marginal or unusual marking and flag these for further review. If the number of flagged issues falls below the vote margin, then there is no way those issues can affect the outcome. In any case, they can be exhaustively reviewed and adjudicated, leaving no doubt as to the outcome.

[10] There is the possibility that in an election that utilizes ballot images could reduce the risk at C1 to 0%, if the ballot image is inspected and verified by the voter during the voting process. Unfortunately, no election systems that now exist offer this feature.

Once ballot images are created and properly secured, they can be made available to the public (at the appropriate time) and to third party auditors. Those parties can implement their own software or crowd-based solution to produce their own audit tabulation to either support or challenge the official result. For any of the ballots where the independent auditors do not agree, these images (and potentially paper ballots) can be further reviewed to determine the accepted result. The auditing software or crowd teams can then use modified voter intent evaluation in the future, allowing all parties to improve.

For the lowest risk, Ballot Image Audits should include image validation to mitigate the possibility of some limited types of hacks that might occur between image creation and securing the images at C2 (while also automatically changing the cast-vote record for those images). For only these few types of hacks, we have to include C5 and C6. Again, hacks affecting only CVR are detectable even without doing image validation.

Image validation simply resolves and constrains C2

C2(image validation) = C5*C6*C12

C12 is determined by sampling risk equation to determine n, the number of samples[11]:

n = CEILING (LOG(risk) / LOG(1-margin/2))

where margin is the pair-wise margin expressed as a fraction, so that 5% is 0.05, and the same is true for risk, the desired risk limit. So ballots are randomly drawn and compared with the image. This increases C2 accordingly, reducing the risk at this stage.

With non-BIA RLAs, the problem is not split like that, and so the calculated risk C12 is correct for all types of hacks, and the other risk components still exist. For BIAs, the problem is split into two types of hacks. Those that affect only the cast vote record, C2 is effectively 100% and those that affect both the images and the cast vote record, C2 is constrained by inspecting the paper.

For BIA image validation, we expect that the ballot images will be identical to the ballots. Sure, there will be some minor differences due to systemic errors, like misfeeds or bad alignments, of perhaps 0.002%. But if we find that any image has been hacked and does not match, (like with a bubble swapped fro the other candidate) then indeed something very significant has occurred. This is different from the statistical RLA approach where a certain number of differences in the cast-vote-record and the paper record can be endured. Here, no non-systemic variation is allowed. If the images are not nearly perfect, then an investigation is warranted.

Administrative procedures can be used to reduce the risk of C2 to fairly low level even if no image validation is performed, particularly if the images are examined and compared in a systematic way during production rather than after-the-fact. This is the way most scanning production teams work in

[11] This is a conservative an approximation based on the assumption that ballots are replaced after each draw. We do not intend for samples to be replaced and so the hypergeometric equation would be more accurate, but much more difficult to solve with virtually no benefit for a large number of ballots. This simpler equation is conservative and sufficient for our purposes here.

corporate America, where scanning of documents is now routine and accepted in court as the original. Those teams use characteristic-based statistical process control[12].

C2 can be further minimized if scanning is separated from interpretation. That is, we would use commercial off the shelf (COTS) scanners that do only one thing, and one thing very well: create images. They do not interpret (parse) the images as many election scanners do today. If the machines are not involved in the interpretation, then the information regarding what the votes are on the ballot being processed is a bit harder to come by, and thus any scheme to influence the election that way must be much more involved. (These scanners tend to be far less expensive as well).

By way of an analogy, eagles have really good eyesight, but they can't read. The section of our brain that does the interpretation of symbols is quite large in comparison and they don't have that part. Same is true in a COTS scanner. They are very good at making images of documents. But without quite a bit of additional smarts, they can't effectively affect the ballot images prior to being secured. Thus, the risk can be minimized probably to a level LESS THAN the risk normally accepted in the sampling RLA audits which are not applied to all contests, and thus in all those contests the risk is not limited. This is in comparison with BIAs, which exhaustively review all ballots in all contests, even if the images are not explicitly validated using sampling. So we have a trade off between sampling RLAs that do not cover all contest and thus are not truly risk limiting in that regard, compared with a ballot image audit, which may not have ballots validated and thus there is a tiny risk that someone might be able to modify the images prior to being secured. It is disingenuous for those promoting the sampling RLAs to point out that the images may not be validated while not disclosing the unfortunate truth that RLA do not limit the risk in local races.

The difficulty to perform a hack to the images can be further increased by insuring that the ballots do not use a fully consistent format for all styles, such as by rotating ballot options.

Furthermore, any optimal hack will require knowledge of the actual (pre-hacked) outcome of the election so a minimum number of ballots can be affected to alter the election avoiding altering it so much that it would cause automatic scrutiny. Non-optimal schemes can be considered systemic, that is, based on a systematic method, like swapping all votes between the two subject candidates or seeking a specific percentage of the vote. Affecting the ballots during the time of C2 is before the actual (unhacked) results are known, and therefore, an optimal hack is not feasible. Any non-optimal hack is easier to detect by definition.

Of course, any scanner should not use any “lossy” compression schemes, and the images must be promptly and properly secured. Using a format like PDF includes internal non-lossy and very efficient compression.

Properly securing ballot images is extremely important, even if a jurisdiction wishes to use sampling RLAs due to C3, the possibility that ballots would be modified after scanning. When ballot images are produced and properly secured, it is infeasible to change the ballot image without detection, and therefore, they are more reliable than the paper itself, except for the risk of C2. Thus, ballot images,

[12] AIIM TR-34 is a standard from 1996 that defines the procedures used to check the fidelity of scanned documents.

properly secured, should be a part of any good audit program. And if the ballot images exist, then why not utilize a 100% Ballot Image Audit instead of the complexity and limited coverage of a sampled audit? It’s a good question.

This importance of BIAs is made clear in contests with few ballots and close margins. In every such case, sampled RLAs will result in costly full hand counts. Also, it is a disappointing fact that most sampled RLA procedures do not include all contests, and if they do, they are not all risk limited to the same level.

BIAs are a type of risk-limiting audit. They do limit the risk, but only if the images are properly produced and promptly secured, and validated. The overall risk calculation is very competitive and they provide for vastly increased transparency as well as competitive audits. Therefore Ballot Image Audits should be the top choice as we move forward.

– Ray Lutz

CitizensOversight.org