Four Fatal Flaws of RLA Audits

Ray Lutz, CitizensOversight.org

raylutz@citizensoversight.org

2019-11-30 V1

2019-12-09 V2 -- Updated details of adding contests not included in proposed CA RLA Regs

URL to web version: https://copswiki.org/Common/M1938

Since 2000, we have had increased scrutiny on elections and their trustworthiness. This has recently boiled up to the top of the national discourse. Any type of audit – if done well – is better than none. But audits done poorly can be worse than nothing. Lately, a great deal of attention has been given to "Risk Limiting Audits" or "RLAs" where auditors review a sample of physical ballots and compare the results with the official outcome. This sounds great, and the theory is sound but when actually applied to real-world elections, we find that RLAs are far from a silver bullet solution.

Popular use of the term "RLA" includes three types: Batch-Comparison Audit, Ballot-Comparison Audit, and Ballot-Polling Audit[1]. All of these approaches are based on statistical samples of ballots which are pulled and interpreted by human auditors, then compared with the official or semi-official results. These audits access directly the paper ballots which have been hand-marked by the voter or at least the voter has had the opportunity to review their votes generated by a ballot marking device (BMD).

Type of RLA	**Description**
Ballot-Polling RLA	Samples are drawn from all ballots in the election, and no computer report (cast vote records file) is required. A margin of victory of the sample is compared with the official margin. For a 5% risk limit, which is the sampling error, it requires thousands of ballots to be sampled for each contest <10% margin. If the contest is not included on all the ballots, the sample may be much larger by that dilution factor.

1 Logically, there is also a batch polling audit but it requires more batches and generally the totals for each batch are available. We use the term RLA because we do not believe these audits are actually that good at limiting the comprehensive risk, and there are other types of audit that can reduce the risk even more. But in popular use as of this writing, RLA includes just the three types.

Ballot-Comparison RLA	All ballots must be individually identifiable and a cast-vote record created for each ballot. Ballot samples are randomly drawn and each ballot is compared with the computer-generated report and discrepancies totaled for each contest. Requires that thousands of ballots be sampled if the margin is <2%. Precinct scanners in use today are incompatible with this method and ballots must be rescanned to create a new cast-vote record and imprint the ballots with an identifier.[2]
Batch-Comparison Audit	Batches of ballots, such as precincts, are hand-tallied and the tallies compared with the computer report. Does not require cast-vote records for each ballot or that each ballot be identified. The auditing process is familiar to election officials who already do hand-tallying in recount procedures. Simplified approaches are best implemented as a fixed number of batches rather than a percentage of the batches, because the detection of hacks is dependent not on the overall number of batches, but on how many are sampled, where auditing 14 batches is enough generally to catch hacks that corrupt as few as 20% of the batches, to a risk limit of less than 5%.

California has historically required the “1% Manual Tally” audit, where 1% of the precincts and Vote-by-Mail (VBM) ballots are randomly audited by precinct or batch. This is a fixed-percentage batch-comparison audit. It is implemented prior to certification and it requires that all contests are audited with at least one batch. Usually batches are either precincts or mixed-precinct VBM batches. Since it does not escalate automatically, it is not a risk-limiting audit, but it is still very useful if conducted properly.

Unfortunately, the number of batches needed for a given risk is not dependent on the total number of batches, but is based on the percentage of batches modified (how well the hack is hidden). If all batches are modified, then auditing just one batch will catch the hack to a confidence of 100%. If we reasonably accept that the audit should be very good at detecting hacks that modify only 20% of the precincts, then this can be detected

[2] ES&S said in an email on 2019-12-09 that their next generation precinct DS200 scanner will be able to imprint a random unique number on each paper ballot so they can be later paired up with the CVR. To use this, it may be better to sample starting with paper first and access the CVR based on imprinted ballot ID number rather than starting with the CVR record because the ballots are not maintained in order in the bin after scanning in any precinct scanner.

with only 14 batches audited to a risk limit of 0.8 ^ 14 = 4.3%, assuming equal sized batches and district-wide contests.

We have carefully investigated how these fixed-percentage batch comparison audits are run in California and how the public can provide useful oversight. Our work started in San Diego County and we (I and other volunteers associated with Citizens Oversight) have provided oversight of these audits for the past 10 years including the top 24 counties in California, the most populous counties in Florida and several other states where audits are used. We were also active in the recount of 2016, especially in Michigan and have reviewed the results of the RLA pilots in Orange County and Rhode Island, as well as the RLA audits as implemented in Colorado.

The fixed-percentage batch-comparison audits as implemented in California are fairly easy to define as they do not have any calculations involved for automatic "escalation", but even without those additional complications, there were fatal shortcuts and "innocent fix-up" which made these sometimes nothing more than theater. Recent changes in the law, due to AB-840, also allowed election officials to exclude sometimes 40% of the ballots from being audited, and that was clearly a step in the wrong direction from a math standpoint.

To get an understanding of how RLAs will pan out when used, we constructed a Monte Carlo simulation, which can simulate thousands of audits with elections of various margins and risk limits, considering the various audit types that were included in the implementation pilots mentioned. This has provided a very thorough understanding of how these audits can pan out.

As a result of this work, we have found that although RLAs can be an effective way to audit elections, there are some serious issues that must be understood as we consider how these should be implemented and draft laws and regulations.

We find there are four serious challenges with RLAs, as popularly defined:

FATAL FLAW #1.
Statistical RLAs Become Infeasible with Tight Margins and are Worthless for "small" Contests with Few Ballots

When margins are relatively large, these approaches work very well indeed because very few ballots are needed to confirm a contest with a wide margin.

On the other hand, for all such sampled approaches, as the margin gets tight, the number of samples required increases, eventually to require a full hand-count.

The Ballot Comparison audit requires the fewest ballots to be scrutinized for any given margin and risk limit, while the ballot polling audit and batch comparison audit require far more. The batch comparison audit processes them in batches instead of one at a time, so it can be more efficient and is the type of audit that many jurisdictions are familiar with.

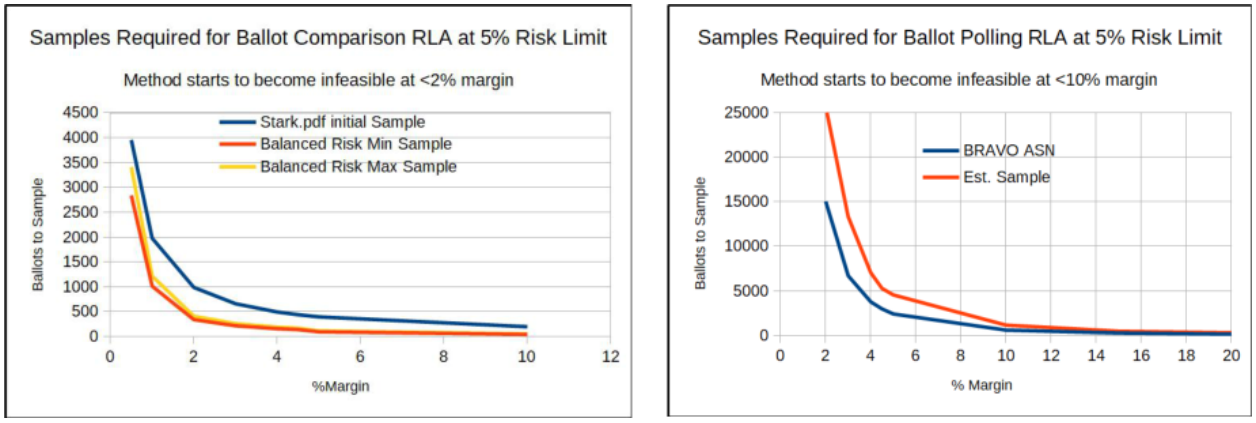
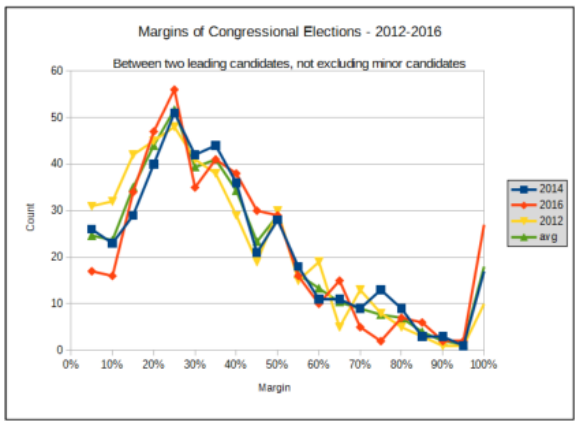


Figure 1: Ballot samples required for Ballot Comparison Audit according to Stark.pdf or balanced risk method becomes infeasible with margins less than about 2%, and Ballot Polling Audit starts to become infeasible at margins less than about 10%, but both of these are very efficient at higher margins.

The good news is that most contests have relatively wide margins (See Figure 2). Recent contests regarding congressional seats nationally show that 90% of those contests have margins over 10%, and typically a contest for a congressional seat has a margin of about 25%. For margins of this magnitude, RLAs perform relatively well. Indeed, sometimes almost too well, as they require very few ballots to be reviewed and this may leave the public concerned that too few ballots were scrutinized.



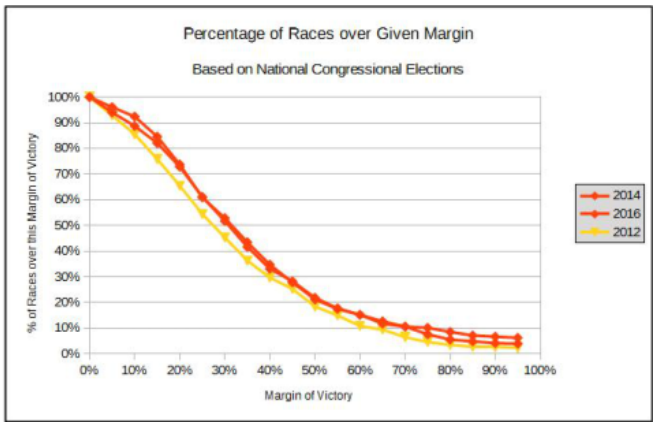


Figure 2: Most Elections have wide margins

But we know that if the contest is a landslide victory, there is probably also very little concern over the results of the contest, while those contests that are very tight, say less than 5%, are the most concern to the public.

Statistically-sampled RLAs start to require a vast number of ballots be scrutinized when margins get tight. Assuming a risk limit of 5%, a ballot comparison RLA becomes infeasible at margins less than about 2% while for ballot polling RLA, they become infeasible at margins less than about 5 to 10%.

If the contest is relatively tight, i.e. less than 10% while also small, auditors may as well just perform a sequential full hand count rather than doing any random sampling that would likely expand into a full hand count because it is less work to just do a sequential and full hand count than start a random audit and then have to switch into a full hand count.

In a recent municipal election in Colorado – which implements ballot comparison RLAs by statute – the margin was so close that they did not attempt to do the RLA at all. Rather, Colorado devolved into reviewing and adjudicating the images, and looking for over-votes and under-votes. Thus, they wound up using a ballot image audit of sorts in the end, but without the other features that should be part of the process to control the risk.

The statistical sampling methods are very powerful and can limit the workload most of the time. But when they fail, we need a way to deal with that without being faced with an insurmountable workload, and then the likely result that the audit is simply terminated without actually confirming the election for these close contests.

Fatal Flaw #2:
Not all races are audited nor risk-limited

Implied in much of the promotional literature about RLAs implies that they will detect flaws in the "election" such that the outcome would differ, and that RLAs will improve voter confidence in the "election." But the reality is such audits can only detect flaws in the contests that are actually audited. Contests that are not audited, do not magically become audited.

One of the foundational technical papers on the topic, "Super-Simple Simultaneous Single-Ballot Risk-Limiting Audits," by Philip B. Stark[3] (S4RLA) suggests that most contests in the election would be audited (underlining added):

> This paper presents some extremely simple methods for conducting the first stage of risk-limiting audits of a collection of contests. The methods allow most contests in an election to be confirmed with a single audit sample of fewer than 1,000 ballots, at a low risk that any of the apparent outcomes differs from the outcome a full hand count would show—unless the audit finds many errors that caused an apparent margin to appear larger than a hand-count margin.

Please note the underlined phrase that *most contests* could be confirmed with a *single audit sample*. In the actual implementation of RLAs, however, very few contests are actually audited. This is not a failure of the theory of RLAs, but in how they are actually implemented.

Knowing that the tightest margin will determine the sample size, auditors reason that by focusing on one key (and likely tightest) contest, other contests will naturally be covered. In Colorado, only one statewide contest is chosen, and one countywide contest in each county[4]. Other statewide and countywide contests can be included in the audit on an “opportunistic” basis, meaning the votes for those contests and they could be evaluated with regard to risk, but will not drive the number of samples. In Colorado, they apparently do collect all the marks from every ballot they sample, but the calculated risk is not reported for any contests other than those explicitly required, and there is no attempt to insure that ballots are sampled from all contests, nor that other contests meet the risk limit criteria. But contests that are not county-wide and include a

---

[3] https://www.usenix.org/legacy/events/evtwote10/tech/full_papers/Stark.pdf

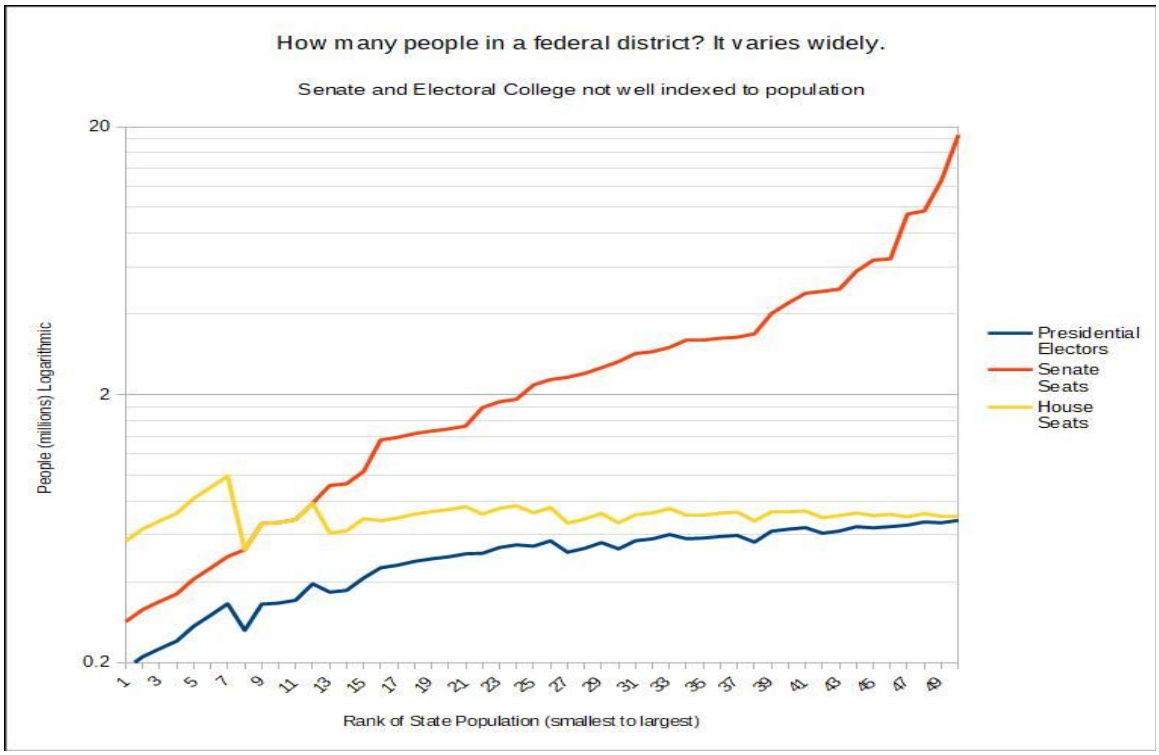
[4] Determined by the Colorado Secretary of State in their RLA regulations.

small subset of the ballots in the county will likely not have sufficient ballots included in the sample to limit the risk to the stated value, like 5%, even if the marks are collected for those contests.

In the proposed CA regulations for RLAs, only three contests are to be audited in each county, generally one statewide contest and two contests either partially or fully contained within the county, selected at random. And although there is no concept of opportunistically expanding the audit to include more contests, the current regulations do propose that a single batch be tallied and compared with the computer report for any contests that is not included in the RLA audit at all[5]. When election officials say "we are doing risk limiting audits" they usually don't mention that these RLA audits provide *extremely poor coverage of the contests,* and promulgate the incorrect notion that by auditing just a few contests, then the results of all contests are reliable.

Randomly selecting the contests to be included in the audits presents two more problems. First, many contests are of low consequence, such as judicial seats (yes/no advisory votes) and contests with only one candidate (who will obviously win), or low-consequence contests like "dog catcher." If low-consequence contests are treated the same as high-consequence contests, like the presidential contest, we get distracted by meaningless audits while potentially ignoring contests that have a higher likelihood of being hacked.

[5] CA Regulations for conducting RLA audits has been proposed by the CA SOS and are open for comment as of this writing. The regulations may change due to feedback. Proposed regulations are at this URL:
https://admin.cdn.sos.ca.gov/regulations/proposed/elections/audits/audits-proposed-regs.pdf



What is spent on campaigning for a given contest can give us an idea of consequence. Seats in the U.S. House of Representatives each include approximately the same number of voters as any other district and we can agree are of relatively high consequence. The average spending on such a seat averages about $1.3 million in the 2018 election. The 2016 presidential elections resulted in spending of about $2.6 billion. That means the presidential contest is about 2,000 times more consequential than a house seat, roughly speaking.

Statewide contests that are not the presidential contest also have various levels of consequence and should *not* be considered equally. For example, a contest for governor is much more consequential than the contest for the insurance commissioner.

Local contests for county-wide positions, such as County Board of Supervisors, are more consequential than dog-catcher in one small city. Yet contests for a mayoral seat in a smaller city in the county may be very consequential and should be subjected to auditing.

Performing random selection of contests where all have equal weights will result in giving too much weight for contests that are inconsequential or of low consequence. At a minimum, contests with only one candidate or advisory votes for judicial seats should

be excluded from the random selection. Then, the remaining seats should be weighted according to consequence, and then according to how tight the margin is. We should audit contests with the tightest margins and of the highest consequence rather than those that are inconsequential and have huge margins, if we have to choose. Random selection is important so that all contests (that are of consequence) may be selected and therefore pose some risk to any election insider who may know which contests are chosen, and just avoid those.

Of course, due to Fatal Flaw #1, election officials would rather audit contests that are a landslide rather than deal with important contests that have tight margins, and for good reason as the audit expands in difficulty and it may be easier to just to a full hand count if the margin is within 2%.

The more significant issue is that randomly choosing contests is far more important than the risk limit applied in just one contest. All of the hoopla over setting the seed randomly pales in comparison to the lack of random selection of the contests, and the very few contests chosen. The proposed CA regulations say that they will audit to a (reasonably tight) risk limit of 5%. But, we must remember that in each county, only three contests are actually audited. The entire risk is increased by the number of contests not audited at all.

By way of example, let's assume that all contests have the same consequence and have the same margin. Thus, there is no difference in which one we choose from that standpoint. If we have 10 contests and we choose one randomly, then we have only a 10% confidence that we will catch an incorrect outcome if our audit process includes no risk of its own. If that contest is audited to a 5% risk limit, then the resulting risk is 90.5% (9.5% confidence), which is extremely low, and does not at all fulfill the notion that the "election has been audited."

There is an unstated and incorrect notion that by auditing just a few contests, then we can rely on the results for all the other contests that were not audited. This is, unfortunately, not the case at all, particularly if the results are targeted by an attack on the central tabulator where the results are changed in one key contest and not any others. This unstated notion is further extended by acting like a 5% risk limit is meaningful when we actually are diluting our confidence in the initial random selection, making the relative low 5% risk limit largely irrelevant.

The election code in California states that the auditing process will provide "comprehensive verification of election outcomes" (underlining added):

> ¶ 15365. The purpose of this article is to provide elections officials with a method to conduct a comprehensive verification of election outcomes through the post-election audit process. This article shall remain in effect only until January 1, 2021, and as of that date is repealed.

Verifying only three contests is hardly comprehensive.

In California, the existing 1% manual tally *does* require that all contests are included in the audit in at least one batch. Additional batches are added after the random draw is completed to make sure all contests are included by adding batches as needed. Batch comparison audits of this kind are sometimes slighted because it is felt that they are too much work and they don't adequately limit the risk. But if we consider that all contests do receive some auditing review, these audits actually limit the risk to much higher standards than the sample-just-a-few-contests approach, even though the few contests audited meet aggressive risk limits.

By way of example, let's assume that an attacker knows that a batch comparison audit is being performed and conceals changes to the tabulation of a given contest to the least number of batches possible, and let's assume further that they can do this by modifying only 20% of the precincts. Taking San Diego as an example with about 1600 precincts (as there were in the 2016 primary), the batch comparison (1% manual tally) audit includes 16 batch samples. For purposes of simplicity, we will assume that all the precincts are exactly the same size, which is not true but makes the math simpler. The chance of detecting the hack by auditing only one randomly selected batch is 20%. But when we choose 16 contests for audit, the chance is much higher. It can be estimated as the chance that we do not choose one of the hacked precincts 16 times in a row, or (80%)^16 = 2.8%. Thus, the risk limit is 2.8%, and we have 97.2% confidence that the hack will be detected (giving the assumptions stated).

Now, let's consider the situation with the new RLA approach where only three contests are included in the audit, and those are subjected to a very aggressive 5% risk limit. In the recent 2018 November election in San Diego, there were 79 contests. If we audit only three of those contests (and for now assuming they are all of equal consequence and so we do not use a weighted sampling procedure), then the chance we would not sample a hacked single contest is (1/79)+(1/78)+(1/77) = 3.84% confidence that the audit would audit the hacked contest, combined with the 5% (95% confidence) due to sampling error, which is almost immaterial at that point, 0.95 * 0.0384 = 3.65% confidence compared with 97.2% confidence in the 1% manual tally approach.

The downfall of Risk Limiting Audits is that they are not well matched to reality. Sure, they are nice in the world of mathematicians who see them as equations, and sometimes argue whether the equation is risk limiting or not. But equations are not the real world. Generally, the RLA techniques consider the entire set of ballots as the population and they sample from that to get to the final results and compare with the outcomes.

But we *do* have other evidence and information that can be used in a more comprehensive review, for example, ballots are always grouped into some sort of batches, either by precinct, random grouping, or some other method. And the election system used to count them has the results for each batch, or it should.

By comparing the results in even one batch with the computer report can detect many types of errors and hacking that might occur. For example, if the columns of the totals spreadsheet has one column per candidate and those column headers are swapped, that's like swapping all the votes in every precinct, and can be claimed to be a clerical mistake. One audited batch will detect that "clerical error."

Another mistake that has occurred from time to time is if the locations of the targets on the ballots are programmed incorrectly. That can swap the votes or just miss one of the targets, resulting in a landslide victory for one option and no votes for the other. One tallied batch per contest will detect this as well.

These can be detected because we do have the totals for each batch (or we should) and those can be used as additional information. The ballot comparison and ballot polling RLA methods disregard that information. That makes it much more difficult to detect the error, requiring far more ballots if you don't use the structure of the batches and the evidence of the batch totals.

What becomes clear is that the sampling of the contests is much more important than the risk limit, and it would be far better to double the number of contests audited at a less aggressive 20% risk limit, giving us 6/79 = 6% overall confidence rather than 3.6%, while making it easier to audit each of the (very few) contests selected. Or better, always audit by batch and add batches to cover every contest, as is done now.

The approach of sampling very few contests for treatment with statistical audited is a very bad idea. This is why this is a fatal flaw of the RLA audits being proposed, even though the math is correct when applied to one contest.

Fatal Flaw #3:
RLAs are infeasible for auditing many small, non-overlapping contests

Even if election officials wanted to implement RLAs so they would cover more contests, the reality is that they are very difficult to apply when there are many contests in small, non-overlapping districts.

Some RLA advocates have asserted that all, or most, contests will be covered by the RLA audits, as described in the prior section. This is even implied in the technical literature and the laws say the audits will be "comprehensive". But covering all contests with a reasonably high level of statistical confidence is very difficult, because many of the contests are not jurisdiction-wide contests, but exist in non-overlapping districts. It isn't possible to audit one or a few contests and then opportunistically expand the audit to cover all other contests. Each contest has to be sampled appropriately, and this results in a vast number of samples required.

For example, all mayoral races do not overlap with each other. City council districts have been commonly broken up into separate districts which do not overlap with each other. Yet, each one of the sets of non-overlapping districts might overlap with some of the districts in the sets of other non-overlapping districts. This results in a large number of ballot styles, where each style has a different set of contests on it, such as exactly which water, school, hospital, fire, and city council districts apply for the voters that receive that style.

In an RLA, there are two ways to select ballots to cover those small districts. One is to choose them randomly of all the ballots in the set. This would typically be the case in a polling audit where there is no organization to the ballots. Assuming the district only includes 10% of the ballots of the entire county, 10x more ballots than the actual number required would need to be pulled because of the dilution factor. Now if the ballots were sorted by precinct, then it would be much easier to choose samples from the set of precincts that include certain contests. But still, the number of samples is related to the margin, and does not decrease just because the district is small. So again, a vast number of samples results.

In fact, the S4RLA document clearly defined the diluted margin and mentioned that the number of ballots sampled is related to the diluted margin:

> The diluted margin μ is the smallest margin in votes among the contests under audit, divided by the total number of ballots cast across all the contests under audit. So, for example, if we are auditing five contests in a jurisdiction where 100,000 ballots were cast in all, and the smallest margin among those five contests is 2,000 votes, the diluted margin is μ = (2,000/100,000)×100% = 2%.
>
> The diluted margin plays an important role in the new procedure: The sample size for the first stage is inversely proportional to the diluted margin.

To more efficiently audit any of those small districts requires that the ballots be sorted at least by the precinct so individual precincts can be selected. If the ballots are individually identified as in a ballot comparison audit, there is a possibility that the ballots could be pulled to cover those contests relatively efficiently. But even in the best of cases, the number of samples required will be multiplied by the number of nonoverlapping districts that are to be audited.

Performing RLAs uniformly and with stated risk limits across all contests, even if you could find the ballots, would be very difficult in districts that have a plethora of local contests. For example, in Dane County, WI, has approximately 340 contests and 206 ballot styles. Any one ballot style presents only a few countywide contests and a few local contests. This is like conducting 206 separate RLAs with the 206 times more samples than one contest (if they all have the same margin).

Sampling the ballots randomly is always more difficult than just performing a sequential full hand tally of that contest batch by batch, and so it is cost effective to just hand-tally that contest when the margin gets below 10% for a polling audit and 2% for a ballot-comparison audit, and this particularly true for small contests. Thus, Risk Limiting Audits become so difficult that they just become full hand tallies, and thus, this is a fatal flaw.

Fatal Flaw #4:
RLAs are complex, difficult and include “Innocent fix-up” hazards

In the ballot polling or ballot comparison audits where individual ballots are sampled, extreme care is required in doing the audit itself. The process of pulling and evaluating the ballot samples, how the data is entered and then compared must be done very carefully, and without making corrections that will defeat the audit itself.

The ballot comparison RLA is particularly difficult because all the paper ballots must be individually identified and organized so every single ballot can be located and matched up with the cast vote record for that ballot. When done in very large districts, this process can become so onerous that it is arguably more difficult than doing the election itself. Certainly, we want the auditing process to be simpler and easier than the process being audited, or we are not gaining any ground. Then we have to audit the auditing process and if that isn't easier, then even with the secondary audit, we are no better off. A full review of all the risks includes more than just the sampling error[6].

What is no doubt the case is that the ballot polling audit and ballot comparison audits are more difficult than the batch comparison audits that are currently in use. Pulling entire batches is easier than randomly pulling all the individual ballots from all batches. Therefore, we recommend that the batch comparison audits be the top method for performing such audits because it reduces the complexity of performing the audit so human error can be reduced.

But human error has another dimension. What we have witnessed in actual election audits is the “innocent fix-up” hazard, where a departure from correct audit procedure defeats the effectiveness of the audit.

During the election canvass, election workers are in the mode of fixing problems and correcting issues that come up on a nearly unending basis. In the audit, however, such corrections are generally not allowed, because it then eliminates the usefulness of the audit results. “Fixing the audit” is not allowed, most of the time. And with a very small sample of ballots in the ballot-comparison audit, the procedures must be absolutely pristine. Such innocent fix-up is virtually impossible to avoid by election workers who are accustomed to fixing problems, are actually auditing their own work, and of course want

[6] See "Comprehensive Risk Estimation of Election Audits" https://copswiki.org/Common/M1913

to produce a clean audit report. But we must emphasize that a clean audit report does not mean the audit is clean.

Carefully designed procedures can reduce the possibility of such innocent fix-up errors, such as separating the review of ballots from the knowledge of the computer results, and mandating that standard hand-marked tally sheets be used. We have proposed such procedures and to a great extent the pilot in Rhode Island and in Orange County did incorporate very good procedures. Yet, we have a great fear that the audits will not provide the sort of check we need and may devolve into nothing more than theater. Election officials will go into their back room and then return to announce that "everything is fine," while observers understand nothing.

Let me give you an actual example. Los Angeles is the largest election district in the nation with 10.6 million residents and about 4,500 precincts. They carefully randomly draw 1% of the precincts (about 45 precincts) in a big affair with 4500 coupons in a large raffle-style drum, which is all fine. Then, they have teams manually tally the ballots in each precinct, and we have no concern with that process.

When they compare than manually tallied result with the computer report, this is the critical point in the process, and where they diverge from best practices.

If the tally matches or nearly matches, they report it without further work. But if the tally does not match the computer report by a significant number (maybe 10 or more, which we are certainly interested in, because it might flag where the cast-vote record was changed maliciously), instead of reporting the result of the manual tally and the discrepancy, they would then take the stack of ballots and rescan them, creating a new computer report. So far, this could be all well and good. The rescan should only be used to diagnose the cause of the error.

But here is where they make the big mistake: They would then only report discrepancies with the new computer report, which would always match perfectly, thereby effectively covering up the original discrepancy. No one really understood that this was a violation of the audit protocol, not even those volunteers that were attempting to oversee it. Fixing a precinct that is bad does not indicate that the audit is clean, but instead should raise a very large flag that something is seriously wrong. Unfortunately, this lack of compliance with careful audit protocol made their audit nothing more than theater (and it continues to be the case).

Evidence-Based Audits with More Evidence

In summary, what we find is that the RLA procedures being promulgated as the "gold standard" of auditing are hardly a good match to actually insure the results are correct, even though the math may look fantastic at first glance. The way to fix this problem is to use "evidence-based" auditing rather than the strict "inspect only paper" RLA approach. We can define evidence based auditing to be audits based on all the evidence available, rather than saying that the only evidence worth checking are the original hand-marked paper ballots, even though we don't want to ignore that important evidence.

RLA advocates rarely admit it, but hand-marked paper ballots are subject to hacking too. In fact, one pen in the hands of an attacker can alter an election by just adding marks on contests where the candidate they don't want already has a vote, so that it is over-voted, or voting for their desired candidate on those ballots the voter left that contest blank. In districts like San Diego, where they allow the use of white-out tape with no logs, reports or a second set of eyes, whiting-out a candidate you don't want is another option.

Modifying the paper ballots can happen prior to the election, as recently detected in North Carolina[7] where absentee ballots were modified prior to being counted, or after they are received, or even to tip the scales very slightly if there is a recount that gets down to a single vote. An RLA audit won't detect these hacks at all.

The primary piece of new evidence we have now, and only available in recent years, is the set of all the ballot images. RLA advocates may say this is untrustworthy, and in the general case it is, but it is a problem that has been tackled in other domains for several decades, as we have moved from paper to electronic documents in all quadrants of our economy. And secretaries of state generally accept electronic versions of legal documents and promote their use, within the "trusted system" concept.

Indeed, the ballot images will defeat any changes to the ballots after they have been scanned, and probably if there was a difference that could be explained as a mark being added, the ballot images would be trusted more than the paper.

7

https://www.nbcnews.com/politics/elections/key-witness-testifies-tampering-absentee-ballots-n-c-house-race-n972896

We have developed some recommendations[8] for securing ballot images so they can be treated with the same "trusted system" concept already used in our legal system today, where documents produced by trusted systems are just as good as the original in business transactions and a court of law. These recommendations for securing ballot images and creating trusted systems does rely to some extent on the software certification process and that the vendor must attest that they have not designed in back doors to modify the images prior to being counted. But we do not rely solely on that evidence.

Indeed, with the paper ballot evidence, which we can routinely sample to insure the images are an authentic reproduction of the ballot (to the extent that the vote can be correctly determined), if there is a difference, the only way it could have happened is due to a back door, and that puts the vendor at risk. If that ever does happen, then it is a simple matter to rescan the ballots get the correct result and then ban the vendor from ever participating in elections again.

An independent auditing service can take those ballot images and create an independent tabulation of the entire election, with precision down to the ballot. We advocate that this should be a standard practice, coupled with a fixed-size batch comparison audit, thereby tallying the paper as well. The ballot image audit will detect all attacks that occur after the ballot images are created, such as the swapped columns and x,y target mistakes, as well as just "change the outcome" attack, as was documented by Bennie Smith as "fraction magic."[9]

At the same time, we firmly believe the best way to vote is with hand-marked paper ballots and not touch-screen machines with internal storage or ballot marking devices. Those may be okay for disabled voters, if we think that is the best solution instead of using those expensive machines is to just have certified helpers who can help voters with disabilities to complete and verify hand-marked ballots.

Paper is important, but we must not disregard the very important new evidence that we now have from all modern voting machines -- the ballot images.

---

[8] "Securing Digital Ballot Images to Enable Auditing" -- https://copswiki.org/Common/M1936 have been submitted to the election cybersecuity working group at NIST.

[9] https://www.youtube.com/watch?v=8ezmpqwVEnM

Some guidelines for a clean audit

Auditing the original paper ballots is an important component in any thorough election audit. We recommend that batch-comparison audits be performed, including all contests (with at least one batch), and using a fixed-size (not fixed percentage) with at least 14 batches. These are better than the proposed RLA procedures being promulgated. To save time and effort, we should turn to using more evidence, and utilize the ballot images in a ballot image audit. Nevertheless, doing any audit properly is still extremely important.

There are a few very important requirements for a clean audit:

1. The computer report ("Cast vote records") must be frozen prior to the selection of contests and the batches to be audited, and published down to the audited unit. If the audit is a batch-comparison audit (such as the kind used in CA), then the report must be published in advance, and broken down by batch, prior to the random draw. (Many districts do not publish the full report for batch comparison audits for the VBM ballots which are not sorted by precinct.) This report must also be frozen prior to the selection of the contests. We believe that including all contests in the audit is required to fulfill the concept that the audits are comprehensive, and all ballots should be included in the audit.

2. Audits should include all contests. But if for some reason fewer contests are chosen, choosing the contests randomly according to their consequence and inversely to the margin of victory is most important. Then, the random selection of batches or ballots can be done by choosing a random seed after all the evidence of the election has been secured, by rolling ten-sided die.

3. The audit team should not have access to the computer report until they have completed their tally process. Otherwise, they may be tempted to seek to arrive at the totals in the report during the tally. A good way to do this is to split any batches in two and have two tally teams tally half, so each cannot seek the result, even if it is published. And rescanning the ballots and using the new computer report must be banned. Only the original and official report can be used. "Fixing" the computer report for a batch that does not match is a violation of protocol.

4. The audit team should use hand-marked paper tally sheets that can be easily scanned and published prior to entry into any auditing software. DRE-like

software which does not have software independence should not be used in the audit process. Indeed, except for performing weighted random selections, RLA auditing software should be no more complex than a spreadsheet.

5. Preferably, the audit team should not be the same people who conducted the election.

6. The audit should be open to public observation, video recording, and the results fully published so it is feasible for any outsider to confirm the results. The act of sampling the ballots and pulling them from storage must also be observable.

Our Recommendation

We recommend that regulations include the following:

1. Election equipment must create ballot images.
2. Ballot images just be properly secured and published.
3. Ballot image audits should be performed by an independent auditing service prior to certification to verify the results based on the ballot images.
4. Districts should use a fixed-size batch comparison audit with approximately 14 batches audited, randomly chosen from all batches including all ballots. Even one substantial difference that is detected should prompt a full review of that contest, including review of the ballot images and the paper ballots.

###

About Ray Lutz

Ray Lutz holds a Master's degree in electronic and computer engineering and has significant industry and standards experience in document processing equipment, including printers, scanners, facsimile, imaging, etc. Also was involved in a test-strategy development group for testing VLSI (very large scale integrated) circuits, and ran a quality assurance department in a manufacturing setting. Founded Citizens Oversight in

2006 and involved in election integrity oversight, particularly of election audits, and mainly with respect to those audits in California.

Contact Information:
raylutz@citizensoversight.org