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
2020-07-25 V3 -- Updated regarding increase of VBM voting during COVID-19 pandemic
2023-08-09 V4 -- Minor update including summary of flaws
2024-01-07 V5 -- Added Appendix regarding Colorado RLA issues from the 2023 election.

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

Summary of the flaws:

1. Statistical RLAs Become Infeasible with Tight Margins and are Worthless for "small" Contests with Few Ballots
2. Not all races are audited or risk-limited
3. RLAs are infeasible for auditing many small, non-overlapping contests
4. RLAs are complex, difficult and include “Innocent fix-up” hazards

Introduction:

Since 2000, we have had increased scrutiny on elections and their trustworthiness. This has recently boiled 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

[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.

audits directly access 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.
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

[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 an 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.

	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[3].

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 with only 14 batches audited to a risk limit of 0.8 ^ 14 = 4.3%, assuming equal sized batches and district-wide contests. Of course, this is only a rough rule of thumb, because it is based on an assumption of how well a hacker could "hide" the changes. But a review of some ballots is always better than none.

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 (other volunteers and I associated with Citizens Oversight) have provided oversight of these audits for the past 10 years. We have also provided oversight for the top 24 counties in California, the most populous counties in Florida and several other states where audits are used. We were 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.

[3] Unfortunately, California allows counties to ignore all ballots that are processed after election day results, including all later-arriving (but timely) vote-by-mail ballots. In 2016, in San Diego, for example, they excluded 285,000 ballots from the 1% manual tally audit. Therefore, this is a fairly significant drawback to the way these audits are implemented.

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, fatal shortcuts and "innocent fix-up" (which will be explained later) can make 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. As vote-by-mail is increased in use due to the COVID-19 pandemic, the ballots omitted from the audit in CA may increase to a majority.[4]

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. Simulations of this type can avoid mathematical mistakes or simplifications and can avoid difficult mathematics that are difficult to solve exactly even for the best statisticians.

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 laws and regulations drafted.

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.

[4] See letter to Gov Newsom on this issue here: https://copswiki.org/Common/M1947

On the other hand, for all such sampled approaches, as the margin gets tight, the number of samples required increases, eventually requiring 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 in terms of accessing the samples, and is the type of audit that many jurisdictions are familiar with, as it is very similar to a hand recount.

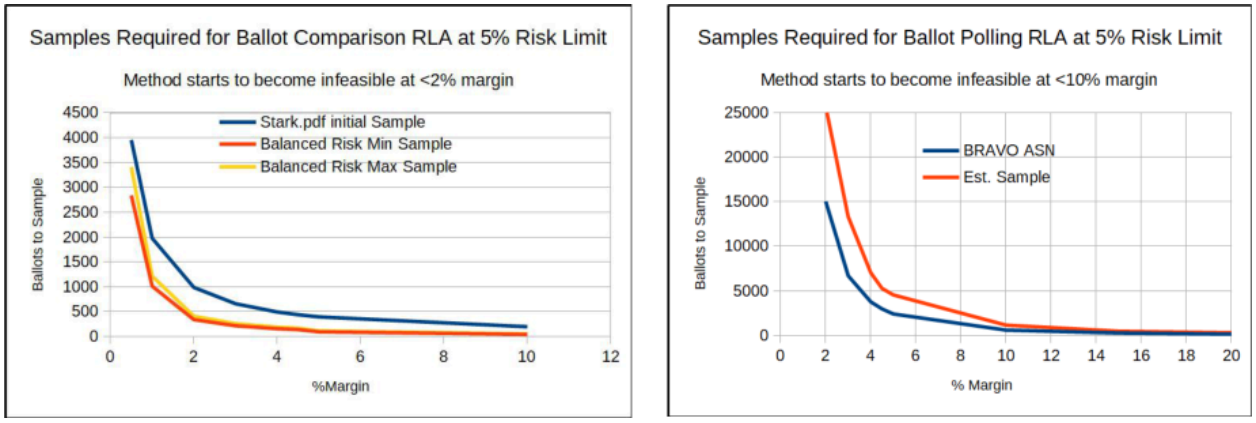
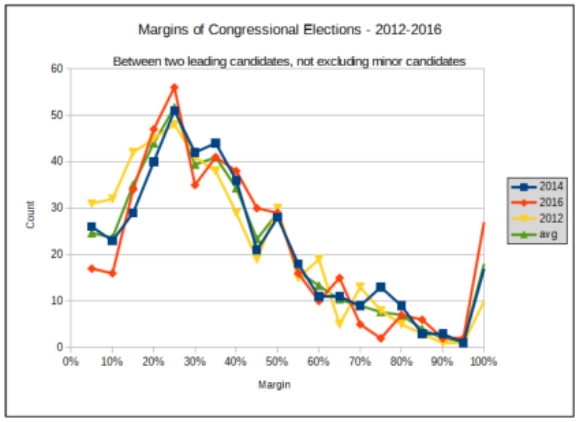


Figure 1: Ballot samples required for Ballot Comparison Audit according to Stark.pdf or balanced risk method becomes infeasible with margins less than about 2%, and Ballot Polling Audit starts to become infeasible at margins less than about 10%, but both of these are very efficient at higher margins.

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.



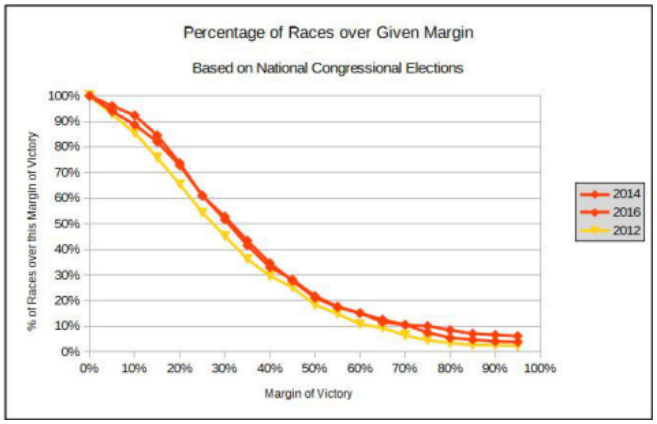


Figure 2: Most Elections have wide margins

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 concerning to the public.

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

In most districts that are relatively small, if the contest margin of victory is relatively small, i.e. less than 10%, if it is also a contest with few ballots, 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. It is less work to just do a sequential and full hand count than start a random audit, throw out those results, and then have to start over and do a full sequential hand count anyway. Batch comparison audits don't have this problem. The batches already processed can be included in a batch-oriented full hand count and can be easily included in the total. The ballot-sampled approaches are simply not efficient when many ballots must be sampled, long before a full hand count is necessary. Switching gears to a sequential full hand count will likely waste all the work already done in the single-ballot sampling mode.

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[5]. Rather, Colorado turned to reviewing and

[5] Need to add the reference to this event.

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. The likely result is that the audit is simply terminated without actually confirming the election for these close contests, exactly what we don't want.

Fatal Flaw #2:
Not all races are audited or risk-limited

Implied in much of the promotional literature about RLAs is 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[6] (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 implemented, because to implement them fully is just a lot more work.

---

[6] https://www.usenix.org/legacy/events/evtwote10/tech/full_papers/Stark.pdf The details of Colorado are provided in Appendix 1.

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[7]. Other statewide and countywide contests can be included in the audit on an "opportunistic" basis, meaning the votes for those contests 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. There is no attempt to ensure 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 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 CA regulations originally proposed for RLAs, only three contests were to be audited in each county, generally one statewide contest and two contests either partially or fully contained within the county, selected at random. (These regulations have since been defined to include "each contest" which we have asserted means also "all contests", which is an improvement in the coverage of the audit, but as a result, will likely become so onerous that no county will attempt it. As of 2020, we now see that they are revising this yet again so that election officials can pick and choose to use the RLA only for some contests.) And although there is no concept of opportunistically expanding the audit to include more contests, the initially proposed regulations provided that a single batch be tallied and compared with the computer report for any contest that is not included in the RLA audit at all[8].

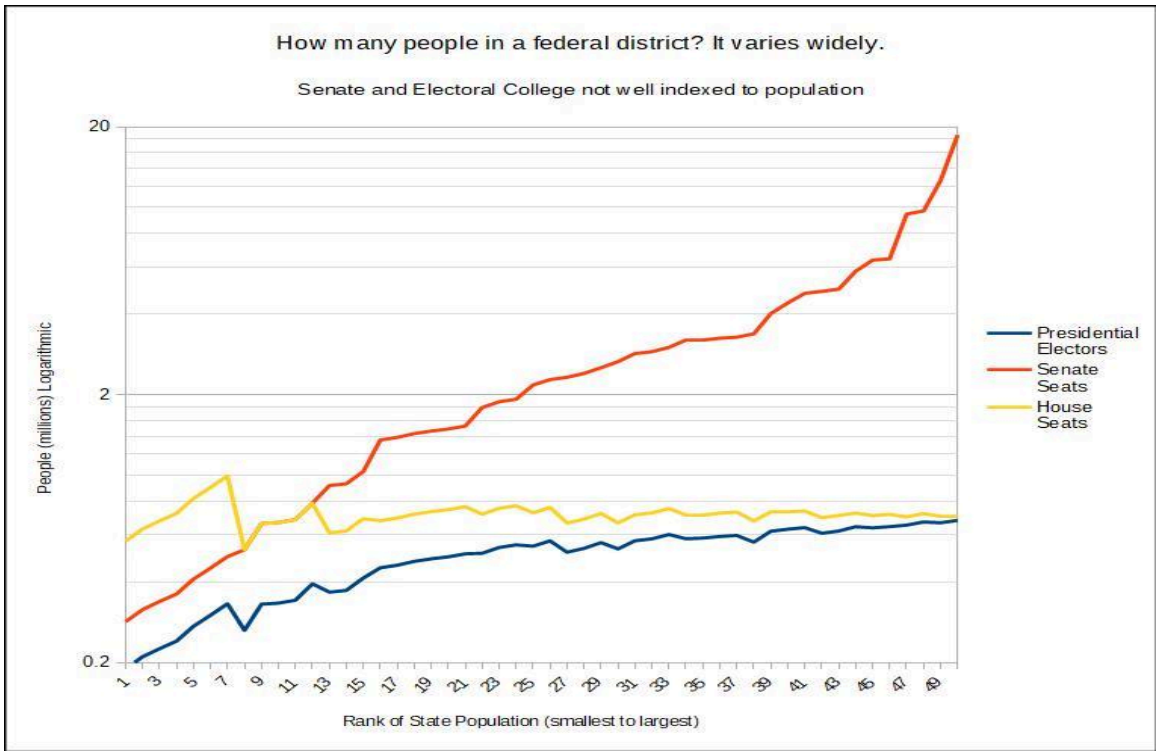
When election officials and RLA advocates say "we are doing risk limiting audits" they usually don't mention that these RLA audits provide *extremely poor coverage of the contests,* and promulgate the incorrect notion that by auditing just a few contests, then the results of *all* contests are reliable.

---

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

[8] 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

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. 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.



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, and using campaign spending as the metric.

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, or an advisory referendum on whether daylight-saving time will be used or not.

Local contests for county-wide positions, such as for seats on the County Board of Supervisors, are more consequential than most other contests, if they are not too lopsided. Also, 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 or given a new-zero weight. 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 any contests (that are of consequence) may be selected. This would pose some risk to any compromised 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 as the number of samples expands, increasing the audit difficulty. It may be easier to just do a sequential full-hand count if the margin is within 2% rather than choosing individual ballots randomly.

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 randomly setting the seed pales in comparison to the lack of random selection of the contests, and that very few contests are chosen. The originally proposed CA regulations would audit to a (reasonably tight) risk limit of 5%. But, we must remember that in each county only three contests would have been audited. The entire risk is increased by the number of contests not audited.

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 assuming 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 to a statistical level of significance."

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 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, and thus the reason they had to back down from these initial regulations which were incredibly weak.

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.

For 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% (because 20% of them were affected by the hack). 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 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), and we assume that the hacker only modified one contest of the 79, then we can calculate the probability of choosing the hacked contest with three drawn as (1/79)+(1/78)+(1/77) = 3.84%. Now, combined with the 5% risk (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 itself 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. A big difference between theory and practice is the fact that pulling samples in theory is error-free, and not subject to any manipulation. Such is not the case in reality.

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

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. Auditing just one 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. This is a configuration mistake that can have a very major impact on the results. 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 only if we have the totals for each batch 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. 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 auditing 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. Also, 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 to 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 from 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, the 2019 Spring election in Dane County, WI had approximately 340 contests and 191 ballot styles. Any one ballot style presents only a few countywide contests and a few local contests. This would be like conducting 191 separate RLAs with then 191 times more samples than one contest (if they all had 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 is 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 and entering and comparing the data must be done very carefully 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[9].

In a recent review of RLA audits in CO, we counted 86% of the discrepancies detected were due to the audit itself, while only 14% were true discrepancies. It appears that the audit process is very difficult and not finding the correct ballot or making audit board mistakes is 6 times more likely than finding any true error.

There is no doubt 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.

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

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-ups are virtually impossible to avoid by election workers who are accustomed to fixing problems, are actually auditing their own work, and, of course, want to produce a clean audit report. But we must emphasize that a clean audit report does not mean the audit is clean, just the opposite is true. If an audit report claims that no discrepancies were found at all, then that is a red flag.

To reduce the possibility of such innocent fix-up errors, carefully designed procedures, such as separating the review of ballots from the knowledge of the computer results and mandating that standard hand-marked tally sheets be used, are recommended. We have proposed such procedures and to a great extent the pilot in Rhode Island and in Orange County did incorporate very effective procedures, but even these would need improvement to avoid this hazard. 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[10]. They carefully randomly draw 1% of the precincts (about 45 precincts) in a big affair with 4500 coupons in a large raffle-style drum. Then, they have teams manually tally the ballots in each precinct. We have no concern with the process to this point.

The critical point in the process and where they diverge from best practices is when they compare the manually-tallied result with the computer report.

If the tally matches or nearly matches, they report it without further work. But if the tally does <u>not</u> match the computer report by a significant number (maybe 10 or more, which we are certainly interested in because it might flag

10 Los Angeles has since moved to a vote-center model, reducing the number of locations to 20% and not sorting ballots by precinct, but processing them in batches.

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. However, 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).

This fatal flaw can be more generally defined as adopting procedures that will make the audits easier and come out "clean" but will not actually implement risk-limiting procedures. We find that in Colorado, the procedures are far from producing results that will actually limit the risk to the stated value for more than just the very few targeted contests. The details of what is done in Colorado is very informative and is provided in Appendix 1.

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 ensure 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 left blank. In at least one district, whiting-out a candidate you

don't want is allowed without logs, reports or another set of eyes (but we think this practice has now been stopped.)

Modifying the paper ballots can happen prior to the election, as recently detected in North Carolina[11]. In this case, absentee ballots were modified prior to being counted, or after they were received. They could 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 tabulation 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.

We have developed some recommendations[12] for securing ballot images so they can be treated with the same "trusted system" concept already used in our legal system today; that is, 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 rely to some extent on the software certification process and that the vendor must attest that they have not designed "backdoors" to modify the images prior to being counted. But we do not rely solely on that evidence.

Indeed, with the paper ballot evidence, we can routinely sample to ensure the images are an authentic reproduction of the ballot (to the extent that the vote can be correctly determined). If there is a difference, it could only have happened due to a backdoor, which puts the vendor at risk. If that ever does

---

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

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

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."[13]

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, but we think the best solution might be to simply have certified helpers who can assist voters with disabilities to complete the ballots and verify the marks are as desired, rather than investing in expensive machines that still have so many hazards.

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.

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

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

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. To fulfill the concept that the audits are comprehensive, all or nearly 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 dice, typically to choose a 20-digit seed.

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. 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, to audit all contests with single-ballot precision.
4. Districts should use a fixed-size batch-comparison RLA 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. If a ballot-image audit is utilized, the contests audited can be reduced to just a few (perhaps three) consequential contests by weighted-random selection.

###

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. He 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. Ray founded Citizens Oversight in 2006 and is heavily involved in election integrity oversight, particularly of election audits mainly in California.

Contact Information:
raylutz@citizensoversight.org

Appendix 1: RLA's don't necessarily limit the risk

Colorado has been on the cutting edge of implementing RLA audits as the first state to enact legislation calling for the use of risk-limiting audits (C.R.S. § 1-7-515) in 2009. The secretary of state selects one state-wide contest and one county contest which "drive" the risk-limiting audit and determines the number of ballots that need to be pulled, even though marks for all contests are reviewed by the audit board.

It is important to realize that since the number of ballots is set by one targeted contest, the risk is appropriately limited only for that one contest. If other contests have tighter margins, or are limited in geographic scope, then the number of ballots required to be reviewed for those tighter contests will need to be higher than what is required by the targeted contest to meet the risk limit. But no attempt to limit the risk in MOST CONTESTS is not performed. Therefore, we can observe that the implementation of RLAs do not necessarily limit the risk in most contests.

In the presentation on December 15 in Colorado at the Bipartisan Election Advisory Commission Meeting, Deputy Secretary of State Christopher Beall presented the contests that could have been chosen in El Paso County as the targeted contest[14]:

The chosen targeted contest was the "City of Colorado Springs Ballot Issue 2A", which had a 4,584 vote margin of victory out of a total of 103,038, votes, or about 4.4% margin of victory. This contest was further diluted to only 2.97% because not all ballots in the county have that contest on them. The sample required for this contest was only 245 ballots.

[14] You can view the video at this link: https://csos.granicus.com/player/clip/417?view_id=1&redirect=true&h=b2d4ac95473007d0db4344a2c71bd4b2 starting at about offset 35 minutes, with the El Paso county details at about offset 55 minutes.

El Paso County – Target Race Selection

154,291 Ballots Cast | # of ballot cards: 1 | 95

3% Risk Limit

County	Contest	Vote For	# of Choices	Lowest Winner	Highest Loser	Contest Margin	Diluted Margin	Risk Limit	# of CVRs to audit	Contest Ballots Cast	# of Counties
El Paso	City of Fountain Council Member - At Large	2	3	1,181	1,003	178	0.12%	3%	6,316	3,875	1
El Paso	City of Manitou Springs Mayor	1	3	895	526	369	0.24%	3%	3,047	1,553	1
El Paso	City of Manitou Springs City Council	3	10	538	441	97	0.06%	3%	11,591	1,553	1
El Paso	Academy School District 20 Board of Directors	2	4	16,180	13,923	2,257	1.46%	3%	498	36,087	1
El Paso	Cheyenne Mountain School District 12 Board of Directors	3	5	3,125	2,715	410	0.27%	3%	2,742	10,000	1
El Paso	Colorado Springs School District 11 Board of Directors	4	10	20,510	18,965	1,545	1.00%	3%	728	51,961	1
El Paso	Fountain-Fort Carson School District 8 Board of Directors	3	4	899	889	10	0.01%	3%	112,432	3,808	1
El Paso	Hanover School District 28 Board of Directors	3	5	122	91	31	0.02%	3%	36,268	355	1
El Paso	Harrison School District No. 2 Board of Directors	2	3	3,154	2,370	784	0.51%	3%	1,434	8,522	1
El Paso	Lewis-Palmer School District 38 Board of Directors - District 2	1	2	5,305	4,869	436	0.28%	3%	2,579	14,213	1
El Paso	Manitou Springs School District 14 Board of Directors	3	4	1,190	960	230	0.15%	3%	4,888	3,303	1
El Paso	Widefield School District 3 Board of Directors	2	5	2,597	2,592	5	0.00%	3%	224,864	9,216	1
El Paso	El Paso County School District 49 Board of Directors - District 2	1	2	2,640	1,941	699	0.45%	3%	1,608	6,234	1
El Paso	El Paso County School District 49 Board of Directors - District 3	1	2	1,794	1,351	443	0.29%	3%	2,538	4,445	1
El Paso	City of Colorado Springs Ballot Issue 2A	1	2	53,811	49,227	4,584	2.97%	3%	245	109,279	1
El Paso	City of Fountain Ballot Issue 2B	1	2	1,466	1,231	235	0.15%	3%	4,784	3,875	1
El Paso	Academy School District 20 Ballot Issue 4A	1	2	19,597	14,685	4,912	3.18%	3%	229	36,087	1
El Paso	Ellicott School District 22 Ballot Issue 4B	1	2	814	208	606	0.39%	3%	1,855	2,883	1
El Paso	El Paso County School District 49 Ballot Issue 4C	1	2	13,350	9,105	4,245	2.75%	3%	265	23,615	1
El Paso	Donald Wescott Fire Protection District Ballot Issue 6A	1	2	1,689	1,310	379	0.25%	3%	2,967	6,886	1
El Paso	Donald Wescott Fire Protection District Northern Subdistrict Ballot Question 6B	1	2	1,866	963	903	0.59%	3%	1,245	5,836	1
El Paso	Flying Horse Metropolitan District No. 2 Ballot Issue 6C	1	2	864	576	288	0.19%	3%	3,904	2,434	1
El Paso	Flying Horse Metropolitan District No. 2 Ballot Issue 6D	1	2	837	608	229	0.15%	3%	4,910	2,434	1
El Paso	Flying Horse Metropolitan District No. 3 Ballot Issue 6E	1	2	150	44	106	0.07%	3%	10,607	2,361	1
El Paso	Flying Horse Metropolitan District No. 3 Ballot Issue 6F	1	2	142	53	89	0.06%	3%	12,633	2,361	1

As we look at the column headed with the title: "# of CVRs to Audit", we can see that there is only one other contest that has a required sample less than this, with 229 samples, and one other which is close, 265 samples. All other contests list far more samples are being required, with one requiring 224,000 samples, which would be more than all the ballots cast in the county (154,000). Thus, except for the one other contest which required 229 samples, none of the other contests would be audited to a statistically significant level, and certainly not to a 3% risk limit.

Beall explained that El Paso is the largest county as well as the county with the highest number of contests. You can see in the column entitled "# of CVRs to Audit" the number of samples required for those contests, and you can see what the options were for selecting the targeted contest. There is within the rules in Colorado, a set of criteria to determine which contest should be selected for the target contests. He said the selection of the "City of Colorado Spring Ballot Issue 2A" was chosen to get a broad spectrum of ballots across the county, with the City of Colorado Springs beng the most populous area within the county, as well as a contest that had a margin that would not be overwhelming or underwhelming, looking for the "goldilocks" number of ballots to pull in that county, understanding the county has a

limited capability and we don't want to overwhelm the paid-volunteers that comprise the audit boards.

Unfortunately, this sweet spot means that except for one other contest, no other contests will have sufficient ballots reviewed to meet the 3% risk limit goal.

In Jefferson County, we received information from Colorado citizens regarding the Arvada City Mayor's Contest. The Arvada Mayor's contest was the closest contest in Jefferson County. It included about 45K ballots out of 206K ballots cast, a dilution factor of about 21%. The margin of victory was very close, about 1%. To achieve a 3% risk limit by randomly sampling county-wide (in Jefferson County), about 5,124 ballots are required, and that is a relatively large number of samples that would need to be individually located and entered by hand.

The SOS selected a different contest to guide the audit, designating that the contest for City of Lakewood Mayor would be used as the guiding contest. It has a much larger margin of victory (18.76%), and as a result, only 179 ballots are needed (as the starting sample) to provide a 3% risk limit.

The selection of Lakewood Mayor may be viewed as a prudent decision to reduce the workload for workers in Jefferson county. But as a result, the sample size is insufficient -- in the case of the Arvada Mayor's contest -- to achieve the claimed risk limit of 3%. In fact, the risk (not confidence) is about 80% with this number of ballots. That's 2,567% greater risk than the risk limit target of 3%.

By pulling samples for only this contest within Arvada City, that would require only 695 samples. The 46 samples pulled in the prior round would not need to be reviewed again, so about 650 additional samples should be reviewed.

We sent a letter to CO Secretary of State Griswold regarding this issue: https://copswiki.org/Common/M2004. They did not improve the number of samples to meet any risk limit for the Aravda City Mayor's contest.

Thus, as we can see in this case, the process of selecting only one contest to use as the target contest for setting the number of samples results in far too few samples to actually accomplish any form of risk limiting for other contests that have tighter margins and may have smaller geographic regions.

Colorado claims[15] that:

> A risk-limiting audit is a post-election audit that gives a statistical level of confidence that the outcome of an election is correct. In other words, after performing a risk-limiting audit, we can say that there is a high probability that the reported winners accurately reflect how voters marked their ballots.

But 20% is not a high probability. The current implementation of RLA audits in Colorado does not result in limiting the risk except for in the targeted contest, and therefore, it is actually not a true risk-limiting audit and cannot make claims about the overall election.

[15] https://www.sos.state.co.us/pubs/elections/RLA/faqs.html