Split the Difference: An Election Audit That Works 'In the Wild'

A Protocol Combining the Best Features of 'Flat' and 'Risk-Limiting' Audits

Jonathan D. Simon, JD

Josh Mitteldorf, PhD

Abstract

With audits increasingly coming into focus as a necessary tool for securing and restoring trust in the U.S. vote counting process, developing a template for an audit process that is both conceptually sound and capable of readily being understood and executed in the field is a critical priority for electoral reformers.

Risk-Limiting Audits (RLA) – which peg the size of the audit to a contest's margin of victory -- have received much positive consideration, and have now been adopted by several states.[1] A number of other states continue to employ "flat" audits, sampling a fixed percentage of ballots or precincts.[2] Many states, however, continue to run elections that are either unaudited or ineffectually audited.[3]

The Split the Difference proposal presented here combines the relative simplicity of a flat audit with the precision and labor-saving features of the RLA.[4] *It achieves this by sampling a fixed percentage (generally 1 percent) of the ballots and pegging the accuracy threshold of the audit---that is, the pass/fail percentage disparity between votecount and audit margins of victory—directly and simply to the votecount margin of victory.* This concept would allow a flexible and effective standard to be written into audit legislation and executed simply and straightforwardly—without increasing reliance on yet more computerization or experts—in the often messy real-world that is Election Night.

Why audits are "in"

There is a clear parallel between the lessons we are learning about computerized voting and the lessons we are learning about living more generally in a "cyber-world." In each case the "gift" seemed to be "free," ours for the taking—convenience, speed, ease, expansion of possibilities. In each case we are learning, in relatively short order (though even the brief delay may well have tragic consequences), the substantial hidden costs: in the case of life in the cyber-world, the great costs associated with theft of information and identity; in the case of computerized voting, the concealed counting of votes in the pitch-dark of a privatized cyberspace, the great costs to democracy associated with theft of elections. Protecting identity and protecting elections are major problems with different potential solutions—but neither one is likely to be doable "on the cheap." And both will require a significant adjustment of behavior and expectations.

[1] See, for example, Colorado, at https://www.denverpost.com/2017/11/22/colorado-election-audit-complete/.

[2] See, for example, Massachusetts, at https://www.sec.state.ma.us/ele/elepostelection/postelectionidx.htm.

[3] Thirty-three states have no audit or inadequate audit provisions, according to the Center for American Progress (https://www.americanprogress.org/issues/democracy/reports/2018/02/12/446336/election-security-50-states/).

[4] For a detailed explication of the RLA protocol, see https://www.eac.gov/assets/1/28/Risk-Limiting%20Audit%20Report%20-%20Final%20.CO.pdf and https://www.stat.berkeley.edu/~stark/Preprints/gentle12.pdf.

Several other nations—including Germany, Ireland, The Netherlands, and most recently Norway[5]—have, in response to concerns about security and fraud, returned to public, observable vote counting in the form of hand-counted paper ballots (HCPB).

Because, however, the United States is not a parliamentary system and because its elections therefore typically involve much longer ballots with more contests to count than in these other nations, HCPB have been viewed by virtually all U.S. election administrators as impractical. The public has long been sold on convenience, speed, and entertainment ("Decision 20XX" as a kind of media extravaganza—a Super Bowl of American politics—dependent upon having results by bedtime in every U.S. mainland time zone), and there are massive inertias impeding such a fundamental change in the Election Night ethos as HCPB would entail.

The question thus arises whether any counting and verification processes short of HCPB might be relied upon to secure and protect U.S. elections and restore both public sovereignty and public trust. The first and most obvious step would be eliminating counting processes that are not subject to any form of verification—essentially the Direct Recording Electronic (DRE) method that involves no paper at all.[6] Such votes, which still represent nearly a quarter of all votes in the U.S., are cast on "touchscreen" computers and can neither be audited nor recounted by any method other than asking the computer to "have another go" and spit out the exact results it presented the first time. DREs are regarded by virtually every cybersecurity analyst as an open invitation to undetectable manipulation and fraud.[7]

Ditching the DREs is obviously necessary but, perhaps less obviously, insufficient. Many have been lulled into the belief that simply having "paper"—capable of being manually audited, or recounted if necessary—is an adequate safeguard. There are two major reasons, however, that this is a false security.

First, recounts are a poor approach to verifying and protecting elections: they are expensive—often beyond what candidates in the financially-depleted post-election period can afford—and burdensome. Beyond these impediments are the chain-of-custody issues that arise when a second count takes place days or weeks after the first count. It is very difficult to secure all ballots and prevent ballot substitution, as several recent recounts featuring torn, unsealed, and misnumbered ballot bags have illustrated. Finally, where either mandatory or elective recounts are pegged to a "close" election within a certain margin of victory—often as low as 0.25%; in California an absurd 0.015% is the threshold for state-funded recounts[8]—it stands as an invitation to manipulators to simply shift enough votes to exceed that margin.

Second, experience has shown that audits are too often weak and toothless, and too often set up with the specific intention of "checking the function of the machines" (that is, do they spit out the same result twice?) rather than verifying election outcomes. The absence of a well-designed audit protocol that can be uniformly applied and made subject to public observation and review is an egregious flaw in our current electoral process and one that must be addressed in any proposals for reform. Without an effective audit process, the "good news" that an election is to be

[5] See https://thebarentsobserver.com/en/life-and-public/2017/09/norwegian-votes-be-counted-manually-fear-election-hacking.

[6] Add to that computerized Ballot-Marking Devices (BMD), that convert voter choices into barcodes that are then read and counted by a scanner. Currently being hawked by several established vendors, such as ES&S, such BMDs somewhat sarcastically produce a "paper ballot," but of course one that no voter will be able to verify as having recorded his or her vote (for the defeat of one such bill, see http://gwmac.com/barcode-voting/).

[7] See https://www.csoonline.com/article/3099165/security/a-hackable-election-5-things-you-need-to-know-about-e-voting-machines.html.

[8] See http://www.sos.ca.gov/elections/statewide-recounts/statewide-recounts-faq/.

conducted “with paper” is really a form of “fake news”: it is rather useless for an election to be *auditable* if it is not in fact effectively *audited*.

What an effective audit does and does not have to do

Before turning to the specifics of this Simplified Audit proposal, it will be useful to set forth guidelines for *any* form of successful election audit. Each of the following is a requirement for an effective, trustworthy electoral audit process:

1. All votes in the venue (state, CD, etc.) must be on paper ballots available for human review.
2. All aspects of the audit—including ballot selection, counting, computation, and posting of results—must be performed observably, in public view.
3. Selection of ballots to be audited must be made by random process.
4. If the audit is designed to sample by *precinct*—that is, to employ a full count of ballots at a certain percentage of a venue’s precincts—then the random selection of precincts must be made at the close of voting, with no prior “telegraphing” of sites to be audited.
5. The audit must be completed on Election Night, before the ballots are stored or removed from the places where cast—including the collection sites for early and mail-in voting.
6. The audit must be conducted by local election administrators and open to observation by, at minimum, one designated representative of each major party and one representative of voters not affiliated with either major party.
7. The audit must be *sensitive* enough to detect outcome-affecting mistabulation of votes and *selective* enough to distinguish between such actionable fraud or error and the minor, non-outcome-affecting counting errors, or “noise,” inevitable in any large-scale counting process.
8. The audit must *specify* an “escalation” process, whereby, if the results of the initial audit disclose votecount-audit disparities exceeding established accuracy parameters, provision is made for expanding the audit, up to and including a full manual count of the ballots.

It should be apparent that the burdens of time and effort are significant, both for those conducting and for those observing the audit process. With the alternatives to the audit approach being either a full hand-counting process or an unverified counting process taking place in the pitch-dark of cyberspace, a fundamental choice must be made about the value of secure and verified elections.

For those who despair that “in the wild”—in the messy, at times chaotic world of Election Night—audits will too often fall short of the standard of perfection seemingly set out in the eight points above, there is a bit of good news. Electoral audits are designed to detect significant and impactful mistabulations—those that jeopardize the “who won?” outcome of elections. But an audit will also have done its job—and done it well—if it manages to *deter* bad actors from interfering with the vote counting process with the aim of altering outcomes.

For this purpose, audits need not be perfect in fact. They need to be strong enough and sound enough to pose a serious and significant *threat* of exposure of any deliberate manipulation of the vote counting process with potential to alter electoral outcomes. Once such an interference is exposed numerically by the audit, it is then up to the investigatory process to pinpoint and prosecute its source. If the audit is basically sound and cannot be readily gamed, his is a risk rational would-be election thieves would not be inclined to take.

By making it part of the risk/reward calculation of any such bad actor, the audit would have served its fundamental purpose of protecting each election from interference and theft—whether by outsiders hacking into the computerized process or by insiders programming mistabulation into the process "at the factory." So, while an audit can't be so slipshod in design or execution as to be subject to compromise by the very same forces that would seek to subvert the democratic process through the mistabulation of votes, if it is designed and conducted in good faith and in adherence to the basic requirements set forth above, it will work as a powerful tool to secure our elections from meddling.

Issues with each of the main audit approaches: RLA and Flat

Election audits can be divided into two major categories: 1) sampling ballots, and 2) sampling precincts.

The ballot sampling approach treats all ballots for a given contest (e.g., statewide, congressional district, state assembly district, etc.), however and wherever cast, as if they were collected in one big bucket, from which a representative (i.e., random) sample is drawn and counted. The results of that count (i.e., the percentages for each candidate in a contest, or Yes/No on a proposition) are then compared with the computed votecounts for that contest and a determination made whether the election "passes."

The precinct sampling approach, instead of treating all ballots as collected in one bucket, considers them as discrete batches (precincts) and chooses a certain percentage of those batches (precincts) at random for auditing. In most cases, when a precinct audit is performed, it functions something like a *spot check*, which is to say that the comparison made is between the computer count and the hand count of all the ballots *in that precinct*.

Although it is possible to treat the random selection of precincts as representative of the jurisdiction (e.g., state, CD) as a whole, and thus compare the aggregate audit tally with the votecount for the entire contest, this approach introduces certain complexities into the sampling process, such as the fact that precinct size must be taken into account in order to generate a precinct sample that is representative of the jurisdiction as a whole. Because a given precinct will often have strong partisan character, a small sampling of precincts, even if random, risks a significant partisan skew.

Audits can further be categorized as "flat" or "risk limiting." In the flat audit a pre-set percentage of either ballots or precincts is sampled; in the risk-limiting audit (RLA), that percentage varies and is determined by the margin of victory in the contest being audited (note that for the purposes of this presentation, only binary elections between two candidates [or yes/no on a proposition] are being considered—with suitable modification, the principles can be applied to multi-candidate contests). With RLAs, the larger the margin of victory, the smaller the sample generally needed to confirm it—and this holds true whether sampling ballots or precincts.

Although the precinct sample—or "spot check"—approach is obviously the more convenient, its efficacy has often been called into question. In theory, what amounts to pulling the ballots out of a scanner storage bin, manually counting them, and comparing that count with the machine tally, should provide a check against machine counting error whether that error is the product of a "glitch" or of fraud.

In practice, there are several problems that may compromise the capacity of the audit to detect fraud. The first involves the selection of precincts, in particular the "cherry-picking" of precincts known to be fraud-free for auditing (e.g., Ohio 2004, where "cheat sheets" were provided to ensure that audit and votecount numbers matched[9]). The second involves the problem of timing—the interval between the selection of precincts and the performance of the audit is a time during which ballot substitution and/or scanner recalibration can be performed. The third involves

[9] See https://www.motherjones.com/media/2005/11/recounting-ohio/.

the fact that precinct spot checks do nothing to verify the aggregation process above the precinct level, so that if the fraud is in the central tabulators it will likely go undetected. As a result, the ballot-sampling approach has come to favored in the design of RLAs.

While there is no question that contest-wide ballot sampling is the stronger and less fallible approach, its logistical challenges should not be underestimated. It is by no means undoable. But it behooves us to be realistic about what it entails—especially if mandated to complete the audit rapidly after the close of voting, before ballot chain of custody issues develop (a choice also has to be made about whether to audit the whole ballot or more selectively—this will make a major difference in the overall burden). Whatever the sampling rate, this approach necessitates having both sampler/counters and observers deployed to every place where ballots are gathered (precincts, mailbags, etc.; it should be noted here that, where the scanners produce "digital ballot images," it may provide an opportunity for off-site auditing, which would greatly facilitate the process, though some have expressed concern about the digital chain-of-custody issues raised by reliance on such ballots images). That's a lot of peoplepower in the real world and a lot to organize and worry about—and also a lot of potential for dispute and conflict, particularly if any unnecessary complexities are introduced into the protocol.

How the new Flat-RLA Split-the-Difference Audit works

The Split-the-Difference (STD) Audit presented here was developed with the aim of simplifying the protocol, reducing confusion and potential conflict in the execution of a ballot-sampling audit, without sacrificing either its sensitivity to fraud or selectivity regarding noise.

It turns out that for the vast majority of contests bearing national significance (statewide, congressional, and even most state legislative races), the *size* for a flat, random sample of ballots is one of the *less* important choices. In fact, for statewide and congressional contests, a 1 percent sample will generally suffice. State legislative contests—depending on the size of the district—may ultimately require a larger sample, but we will see that this "escalation" process is built into the STD Audit design.

Of far greater importance than sample size is the selectivity/sensitivity issue—or knowing when an audit result is a red flag that requires an escalated sample, full recount, or some such investigation. This is where bringing in the "risk-limiting" concept works. To conduct an RLA (as proposed by Berkeley Professor Phillip Stark and others) generally requires a recursive algorithm to determine the percentage of ballots to be sampled. In essence, you keep going until the count of sampled ballots guarantees accuracy of the initial (computer) count to within an acceptable and pre-determined level of confidence.[10] As a result, the percentage of ballots to be counted will vary and a separate sampling be necessary for each contest on the ballot, leading to complexities and potential problems in execution. It is fair, based on observation and experience, to say that Election Night administrators and workers want *routines*—the less variation, the better.

The STD Audit developed from the insight that you can conduct a flat audit and build the risk-limiting concept into the escalation provision. *The way to do this is to peg the "accuracy threshold" of the audit -- that is, the acceptable percentage disparity between votecount and audit margins of victory -- to the votecount margin of victory (VMOV).*

The audit legislation can provide, for example, that if the audit margin of victory (AMOV) is *less than half* the VMOV, that will trigger escalation. Note that we do not care about *any* situations in which the AMOV is *greater* than the VMOV—we're concerned here (in simple binary elections) strictly about who won, not by how much.

[10] See note 4.

I cannot emphasize this point too strongly. With apologies to any "purists" in the room, pragmatism must prevail in this enterprise. Given the demands, the delays, the administrative, institutional and cultural resistance—in short, all the inertias, rational or otherwise, that have thus far locked in a manifestly vulnerable computerized vote counting process—the whole quest for reliable audits is inevitably going to be seen (especially by those upon whom the demands are being made) as a bridge too far. It is imperative not to make it an inch longer than it absolutely has to be.

Probably the best way to show the STD Audit at work is through two sets of basic examples.

Our first set will audit a run-of-the-mill noncompetitive election. The vast majority of binary (two-candidate or yes/no proposition) races are not close. While reversing the outcome of such elections might risk failing the smell test, it is not beyond the pale from a technological standpoint. All elections deserve quantitative protection, but if the huge swath of noncompetitive ones can be handled simply and quickly, the audit process as a whole will be seen to be far more practical.

For our noncompetitive example, let's posit a votecount of 70 percent for "A" and 30 percent for "B," a VMOV of 40. We assume that the "one-half VMOV" standard (VMOV/2) has been written into the audit legislation; in this case VMOV/2 = 20. Let's now look at a few possible audit outcomes:

1) If the audit result is 66%A - 34%B, the AMOV = 32; 32 > 20, so do nothing further; the election PASSES.

2) If the audit result is 95%A - 5%B, the AMOV = 90; 90 > 20, so do nothing further; the election PASSES. Note that in this example, the audit is "way off," but not in the direction that would suggest any possibility of an outcome reversal; as a general rule, whenever AMOV > VMOV, the election passes.

3) If the audit result is 58%A-42%B, the AMOV = 16; 16 **<** 20; the election FAILS, so *ESCALATE*; this would be a red flag if, as we began by assuming, the legislation sets the "one-half VMOV" standard. Note that most contests will be the 70%-30% species (i.e., noncompetitive) and in these contests the STD Audit will virtually *never* have to escalate unless either something has gone seriously blooey or some election thief has gone seriously crazy.

It gets more interesting in the instances when contests are close. For our second set, let us take a *competitive* election in which the votecount is 52 percent for "A" and 48 percent for "B," a VMOV of 4. We again assume that the "one-half VMOV" standard has been written into the legislation; in this case, VMOV/2 = 2. Again, let's look at a few possible audit outcomes:

1) If the audit result is 55%A - 45%B, the AMOV = 10; 10 > 2, so do nothing further, the election PASSES (remember, *whatever* the VMOV is, if AMOV > VMOV, the election passes).

2) If the audit result is 49%A - 51%B, the AMOV = -2 (negative because it is now A's margin of *loss*); -2 < 2; the election FAILS, so *ESCALATE*; it can be seen that the impact of our simple formula for escalation is that *whenever* the audit produces a different victor from the votecount, AMOV will perforce be less than VMOV/2, and it will be a red flag triggering escalation by provision of the legislation.

3) If the audit result is 51.2%A - 48.8%B, the AMOV = 2.4; 2.4 > 2, so do nothing further; the election PASSES.

4) If the audit result is 50.8%A - 49.2%B, the AMOV = 1.6; 1.6 < 2; the election FAILS, so *ESCALATE*.

I hope those examples make clear how this audit and pre-set standard would work in practice. It is based on a simple "flat" sampling of a fixed and pre-determined percentage of the cast votes for a given race. This means that, unlike with the standard RLA protocol, ballots don't have to be re-sampled at different rates for each race to be audited. *A single sampling of the ballots will enable auditing of all races subject to audit.*

Note also that the escalation decision was in all cases *independent* of the statistical margin of error (MOE) of the audit. The MOE is a gauge of the likely accuracy of a sampling, with smaller MOEs indicating a more “powerful” sample. Logically the MOE decreases with increasing sample size—but, perhaps counterintuitively, it is virtually independent of the size of the pot being sampled, once that pot reaches about 20,000 in number, as is the case with virtually every election bearing national significance. For illustrative purposes, a random sampling of 1,500 ballots from such a large pot will yield a result within 3 percent of the full votecount 95 percent of the time; a 3,000-ballot random sample will come out within 2 percent; and a 10,000-ballot random sample within 1 percent.

In our STD Audit, however, assuming the sample of ballots is of a reasonable size—which would be provided in nearly all cases by a 1 percent level of sampling—the MOE can effectively be (purists, cover your eyes) *disregarded*. This is because the MOE doesn't really come into play in *noncompetitive* elections—the audit results can be well outside the MOE without triggering the "one-half VMOV" standard, and will work just fine. And in very *close* elections, we will want to escalate if the one-half VMOV standard is triggered, *even if the results are within the audit's statistical MOE*.

Note that, for the purposes of illustration, we assumed a one-half VMOV (VMOV/2) standard for escalation—and that one-half standard gives the new audit its name, Split-the-Difference. That standard could, however, be legislatively established to be higher or lower: for example, a “one-quarter VMOV” standard, which would make escalation less likely and “weaken” the audit; or a “three-quarters VMOV” standard, which would make escalation more likely and “strengthen” the audit. I propose the one-half VMOV standard as being just about right in balancing selectivity and sensitivity.

And, of course, a more *complex* standard could be concocted, incorporating the MOE as a second trigger and essentially hybridizing the standard. But I don't see much value in that. That hands it all back over to the computers and the experts without an appreciable gain in auditing power, selectivity, or sensitivity. The closer to KISS this can be, without sacrificing detective or deterrent power, the better.

Sheep may safely graze

It may be of value, for anyone yet unclear about how the STD Audit is designed, to consider a pictorial/schematic representation of the audit at work. Picture a meadow with sheep grazing that ends in a cliff. There are sheep all over the meadow, some closer to the cliff's edge, some farther away. Each represents a computerized vote-count of one of the contests on a ballot. Then the sheep begin to move. Some move closer to the cliff, some move further away. Their new positions represent the results for each contest given by the STD Audit. The cliff, of course, is the 50-50 line, the place at which there would be a different winner.

We don't have to worry about any sheep that move further away from the cliff (statistically, that will be half the audit results); they're safe. And we don't have to worry about any sheep that move towards the cliff but don't get halfway there (while you don't get a statistical guarantee, the likelihood that the audit is signaling a miscount severe enough to fall off the cliff and change the winner is very low).

But any sheep that moves *more* than halfway to the cliff is in danger of falling off—a red flag, so you escalate to a full manual count (or an expanded count). There will be, unless fraud is rampant and bold, very few, if any, of those. Very close races will, of course, be more likely to escalate—which is how it should be, and just what is provided for in the RLA approach.

The STD protocol is basically measuring via audit how close either riggers or errors have come to the cliff's edge of altering the winner. And it does that with a minimum of sampling/counting sweat, complexity, and opportunity for confusion and error—and no nontransparent calculations whatsoever.

Conclusion

The Split-the-Difference audit combines the best features of the flat and risk-limiting species. The “flat” (predetermined sampling percentage) aspect allows for simplicity of execution, a single sampling for multiple contests, and the development of a routine that is replicable from one venue and one election to the next. The RLA aspect bestows greater efficiency and calls for more labor *only where such labor is truly neede*d *to verify winners and losers of elections.*

The STD Audit is appropriately sensitive to potentially outcome-changing fraud or error, while sufficiently selective to make quick work of elections that are not in such jeopardy. It provides powerful anti-fraud deterrence while keeping Election Night complexity, confusion, conflict, and exhaustion to acceptable minimums. It should thus commend itself to election administrators concerned that the implementation of a serious audit protocol (particularly one of the RLA type) will trigger such issues in their real world—“in the wild,” as it were. It also should gain the approval of election integrity advocates seeking an executable audit with built-in escalation provisions—in short, an audit with sharp teeth and very little fat.

Although the political and administrative inertias, even in "good" places, remain formidable, one can sense, in these politically-fraught and security-challenged times, a certain reweighting of concerns and desiderata—and, with that, the growing prospect of rescuing our elections from cyberspace and the black box in which they have become trapped.

Jonathan Simon is a 16-year election integrity veteran and author of CODE RED: Computerized Elections and the War on American Democracy*. He lives near Santa Cruz, California; blogs at www.CodeRed2014.com/blog, and is reachable at verifiedvote2004@aol.com. He invites comments on this proposal and discussion of all issues bearing on election integrity in the United States.*