

Thanks, responses, and corrections
to the USEnix
“2009 Electronic Voting Technology Workshop /
Workshop on Trustworthy Elections” Reviewers
who rejected the article

Checking Election Outcome Accuracy

Post-Election Auditing Sampling

<http://arxiv.org/abs/0907.3166v1>

My request to the EVOTE/WOTE conference Chairs asking them to forward my comments to the Reviewers was denied, citing a desire to avoid setting new precedence. Fair enough. Hence these comments are publicly posted here:

<http://ElectionMathematics.org/em-audits/US/EVT-ReviewerResponse.pdf>

and I welcome the Reviewers to read them and to post their anonymous responses there.

The latest version of

Checking Election Outcome Accuracy -- Post-Election Auditing Sampling is posted on-line at:

<http://electionmathematics.org/em-audits/US/PEAuditSamplingMethods.pdf>

or at <http://arxiv.org/abs/0907.3166v1>

[EVT/WOTE '09] Paper #7 "Checking the Accuracy of Election..." Not Accepted
Tue, May 26, 2009 at 4:38 AM

-- EVT/WOTE '09 Chairs

David Jefferson

Lawrence Livermore National Laboratory

Tal Moran

Harvard University

Joseph Lorenzo Hall

UC Berkeley/Princeton

RESPONSE TO COMMENTS OF REVIEWER #1

REVIEWER #1: "This paper provides some terminology and background on post-election auditing, including comparing fixed-rate and risk-limiting audits. The paper reformulates or calculates new bounds for number of audit units to select under various circumstances."

ME: Yes true.

REVIEWER #1: "Overall, it felt like the author attempted to cram too much into a single paper, making the paper less coherent and effective than would be desirable. Focusing on a single, novel idea would have made the paper far stronger.

The paper also could have used a far greater number of examples and more frequent reminders of the variables. Often, the formulas were dense and difficult-to-follow because the paper didn't provide adequate intuition. In addition, variables (such as E) are sometimes used with insufficient explanation."

ME: Thank you for this suggestion. I did not realize that my paper was difficult to follow until I reviewed it with this in mind. I had thought that my "Table of Variables" that preceded the use of any variables was made it crystal clear. This Reviewer is absolutely correct that the article needed more running explanations of the variables to make it easier to read.

REVIEWER #1: "The author should consider providing greater justification for using estimates and approximations in some cases. While it is not necessarily unacceptable to use these, rationale should be provided for their use."

ME: Thank you for this comment. All post-election audit (PEA) sample size calculations depend on inputs that are estimates (such as estimates for the number of miscounted audit units that could cause an incorrect outcome or estimates for the amount of maximum possible within audit unit margin error) and so all PEA sample sizes are consequently estimates, albeit hopefully accurate and conservative ones.

The exact error amounts in the initial election results and the exact number of audit units that cause an incorrect outcome can only be determined by an accurate 100% manual count.

Some estimates used to calculate sample sizes are less accurate than other estimates because they fail to employ the precise available data or fail to use precise calculation methods, including using the "2sv" estimate method (where "2s" is the maximum percentage of undetectable error usually $2 \times 0.20\% = 40\%$ and "v" is the number of votes cast) or because they use an arbitrary fixed "error tolerance" ignoring the exact upper margin error bounds (b+w-l) that were recommended for use first in December 2007 by myself and later recommended by Stark.

When the data, tools, or expertise required for the most precise estimates is unavailable then rough estimates may be needed.

I have added an explanation to alert readers that all post-election audit sample size calculations are estimates.

REVIEWER #1: "As a contribution, the author indicates that the paper unifies methods by Aslam et al. and Calandrino et al.,..."

ME: Great. I was hoping that the Reviewers would catch this error. This was a hasty mistake I made in my over-enthusiasm just after discovering the new error bound formulas for all candidates for weighting random selections of audit units that apply to both the Aslam et al. and the Calandrino et al. methods. You are right that these methods are *not* equivalent. I have already included a separate section on the improved PPMEBWR method.

REVIEWER #1: "... but I believe that those papers already reference each other and draw parallels in their mathematical underpinnings."

ME: Yes True. However, both the PPMEB and PPMEBWR methods originally proposed by the Aslam et al. and the Calandrino et al. groups need significant improvements to their sampling weights and sample size calculations in order to assure that they use adequate sample sizes to achieve the desired probabilities for detecting incorrect initial election outcomes.

REVIEWER #1: "In addition, the author appears to be flawed in her critique of one or both of these papers in Appendix D. Calandrino et al. describe a threat scenario in which attackers change a ballot as part of an attack that adds b_1 incorrect votes. Although Appendix D critiques that b_1 votes cannot be added to a single ballot, the attack that Calandrino et al. describe involves adding a single vote to each of b_1 ballots (not b_1 votes to a single ballot)."

ME: That may have been what Calindrino et al. *meant* to say, but I quoted their paper exactly (in Appendix D) including such obviously incorrect statements as:

If C_s is non-null [there is a winning candidate voted for on the ballot], then we need to audit this ballot with probability at least $1 - (1-c)^{(1/b_1)}$, where $b_1 = v_s - v_{(k+1)}$ [the total "reported vote totals" for that winning candidate and just-losing candidate]. Intuitively, one possible result-changing scenario involving an error in this ballot would be to add $v_s - v_{(k+1)}$ incorrect votes for candidate s .

If the Calandrino et al. mathematics were logically correct, it was not possible to determine that from from their mathematics or prose. My article improved the PPMEB method's sampling probabilities to provide a logically coherent method for sampling in cases not only when individual ballots are the audit units but also when larger audit units are used.

REVIEWER #1: "At quick glance, I am not sure about the critique of Aslam et al. either (it may be a simple terminology issue, the critique was a bit unclear)."

Me: The problems with PPMEBWR method originally proposed by Aslam et al. stem from the imprecise and sometimes impossible margin error estimates they used that have the effect of causing insufficient sample sizes to be calculated that do not obtain the desired minimum probability for detecting incorrect election outcomes. There was also a small, but crucial error in the Aslam et al. algorithm for determining the number of draws as discussed in their PPMEBR method by mixing up various winning-losing candidate pairs.

REVIEWER #1: "Minor issues:

Page 3, "Figure 1 shows that 2%..." I believe this should be 3%.

ME: Another good catch. Thank you for catching this so I could correct it.

REVIEWER #1: Figure 3 is difficult to read. "

ME: Thank you. I am a LaTeX newbie who is still trying to learn how to create legible PS and PDF images that display clearly. The current version of my article is improved but not perfect.

Note: Reviewer #1 made some excellent catches, but got it wrong when he claimed that the Aslam et al. and Calandrino et al. methods do not need any improvements.

RESPONSE TO COMMENTS OF REVIEWER #2

REVIEWER #2: "The paper provides some recipes for choosing "how many audits" to conduct post-election in order to provide acceptably high levels of confidence in the overall vote tally. "

ME: Almost. This paper provides recipes "for choosing how many audit units to randomly sample and manually count in order to ensure a desired high level of confidence in an election outcome (the winner).

REVIEWER #2: "The raw statistical results and/or formulas are very well known and can be lifted almost directly from a good statistics text book. (See, for example, "Mathematical Statistics: Basic Ideas and Selected Topics," by Bickel and Doksum.) "

ME: If this statement were true, that would mean that Aslam et al., Calandrino et al., myself and many other authors have been wasting a lot of hours of research developing new formulas to apply to election audits and that despite our training and graduate degrees in mathematics never noticed the formulas you mention. I have also discussed the mathematics of election auditing with several financial auditors who admitted that nothing in their field applied to elections. I doubt the truth of this claim.

Reviewer #2: "... where there is perhaps room for an election specific treatise of this subject would be in documenting, as well as empirically justifying a precise threat model. Unfortunately, the paper is weak here. Some discussions of threat mechanisms are scattered through the paper, but they need to be tightly unified so that a reader can compare results against the model. "

ME: A "threat model" is not necessary for calculating post-election audit sample sizes. In post-election auditing the source of the error is a tangential concern and more key are:

- 1. Is the voting system independently and conveniently verifiably auditable?**
- 2. Can we detect any and all material errors in the counts, regardless of the cause of the errors?**

In other words, the causes of vote miscount are immaterial to the initial sample size calculations for post-election audits because *any* error that could materially affect

the outcome of the election needs to be detected, regardless of its source. A threat model is useful in designing security and in investigating the cause of discrepancies found in a post-election audit, but is not pertinent to sample size calculations. In another paper on PEA procedures I explain why certain auditing procedures are necessary to detect certain threats.

REVIEWER #2: "...moreover, to the extent that I can surmise the intended model, I believe it is incomplete. For example, the author asserts that a "manual recount" provides "100% assurance." but this is clearly an overstatement. Human counters are subject to errors too, and we have seen this in several high profile elections during the last decade. "

ME: Thanks for bringing my attention to the fact that I had neglected to explicitly state my assumptions.

My article now:

- **mentions that post-election auditing *procedures* are discussed in another paper in the same series "Checking the Accuracy of Election Outcomes" and that, as recommended by Luther Weeks of CTVoterscount.org, when differences are found, at a minimum, recounts should be performed by local officials until two counts agree - either the machine and a manual count or two manual counts.¹ and**
- **points out that I assume for the purpose of calculating sample sizes that good manual counting procedures are used that ensure that the manual counts are 100% accurate such as the manual counting methods employed by the Secretary of State's Office of New Hampshire (and unfortunately not used much elsewhere.)**

REVIEWER #2: "It is important to make election officials continually aware of the general issues that are presented. However, the paper contains too much detail for that audience, and many papers on the subject have already been published. "

ME: "Many papers on the subject" of post-election audits (PEA) that "have already been published" contain crucial flaws that must be improved if the post-election audits are to reliably achieve their stated goal of detecting incorrect election outcomes. My article improves upon the prior recommendations made both by myself and other authors and provides an understandable overview of methods to calculate PEA sample sizes.

Note: Does Reviewer #2 think that post-election auditing is of no import?

RESPONSE TO COMMENTS OF REVIEWER #3

REVIEWER #3: "This paper contains some solid content, but aside from many small writing errors, it has several basic problems."

ME: Grammar or formatting errors? I am prone to both.

¹ Results of Post-Election Audit of the May 4th Municipal Election
<http://www.ctvoterscount.org/?p=2077>

REVIEWER #3: "The intended audience for the paper is unclear. For instance, roughly half a page (pp. 7-8) is devoted to proving that of all winning-losing candidate pairs, the just-winning and just-losing pair will yield the largest error bound (upon certain assumptions) and therefore the most conservative sample size (upon other unstated but reasonable assumptions). If one grants the assumptions, this point seems sufficiently intuitive not to merit so much space in the main text. "

ME: Apparently this was not intuitive to other authors such as Stark and Aslam et al. who mixed up various winning-losing candidate pairs in some of their calculations of PEA sample sizes in their prior papers. These incorrect recommendations are being adopted today in some states. However, I agree with you, so I moved this little proof to an appendix.

REVIEWER #3: "However, based on footnote 11, Dopp appears to offer it as a critique of Stark's 2008 CAST paper, which advocates computing and summing the maximum pairwise error bound for each audit unit. All else equal, Stark's method will yield more (or no less) conservative error bounds than Dopp's."

ME: Stark's methods are *not* more conservative. Some of Stark's methods used a fixed small acceptable error tolerance that causes him to in effect ignore the within audit unit upper margin error bounds. Other of Stark's methods cancel any maximum level of undetectability from two sides of an inequality or from top and bottom of a ratio before his final tests or sampling calculations. Stark's mathematical algorithms need improvements.

REVIEWER #3: "Is this a bug or a feature? From Dopp's analysis, one cannot tell, because Stark's CAST protocol incorporates rules for determining when the audit can stop -- an issue that Dopp declares beyond the scope of her paper -!"

ME: First,

Some of Stark's recommendations for "determining when the audit can stop" have been inconsistent with fundamental premises made in calculating sample sizes for risk-limiting post-election audits. The only correct conclusion that may be drawn from a too-small PEA sample size is that the sample size is too small and the audit must be expanded – even when no discrepancies are found. Stark replicated several formulas that others' developed before him, perhaps without noticing it (or at least without citations) but has often applied these formulas in ways that are less precisely accurate than the prior authors.

Second,

I have virtually completed another article on "Post-Election Audit Discrepancy Analysis" that shows an algorithm for making the decision on whether or not to certify the election outcome or to expand the sample size, using simpler and improved that correctly reflect the underlying premises of PEA sample size designs.

REVIEWER #3: "- and those rules rely on his error bounds. "

ME: Whose error bounds?!*!

I was the first author to recommend using those error bounds that you erroneously call "Stark's error bounds", although Aslam et al. also derived the same error bound expression in an intermediate calculation well prior to Stark's employing

them, but then Aslam et al. did not recommend using them and instead recommended the earlier, less accurate \$2sv\$ method that produces inadequate sample sizes in most cases.

Stark's failure to mention that some of the expressions and formulas that he recommends were derived and recommended first before him by other authors is not a reason to name those methods after him.

Perhaps these error bounds should be called "Dopp's error bounds" or "Aslam's error bounds" because Aslam, Popa, Rivest and myself all derived and mentioned the error bounds that you attribute to Stark, well before Stark employed them.

REVIEWER #3: "So, here Dopp wanders into a subject unlikely to benefit a non-technical reader, in a manner unlikely to benefit a technical reader. Other passages pose similar problems. "

ME: Thank you. I'll move that discussion to an appendix.

REVIEWER #3: "The issue of scope seems fundamental. At this point in mid 2009, a paper on audit sample sizes that has nothing to say about when a risk-limiting audit can stop is facially unlikely to make a contribution beyond previous literature, unless it is unprecedentedly accessible to a general audience or focuses on one or more useful technical innovations. "

ME: This parrots Stark's claims. This statement is only true if you want the public to use inadequate initial post-election audit sample sizes that will not reliably achieve their stated probability for detecting incorrect election outcomes or if you expect the public to use methods for analyzing the discrepancies that are based on insufficient initial sample sizes. Yes, the discrepancies must be analyzed and a decision-rule or algorithm is needed, but the first step in conducting efficient, effective scientific audits is to ensure a sufficient sample size that can detect an incorrect outcome in even a very close margin contest that was caused by well-hidden vote fraud. Risk-limiting auditing in elections is not the same as in manufacturing parts or in financial transactions. In efficient, effective risk-limiting election auditing sampling must be related to specific election contest results.

I have addressed the issue of analyzing the discrepancies in a separate virtually complete article.

REVIEWER #3: "MLU (maximum level of undetectability)

Dopp employs a hybrid of Saltman's "maximum level of undetectability" (MLU) and Stark's conservative error bound,..."

ME:

First, "Stark's conservative error bound" was recommended first by myself in December 2007 and was derived first by both myself and by Aslam et al. and not first by Stark.

Again, Stark's method is *not* more conservative, because his methods negate the use of the margin error bounds by employing a fixed small tolerable error level or by introducing calculations where the maximum level of undetectability would cancel if he used one.

Saltman's "maximum level of undetectability" (MLU) is used in conjunction with ***ALL*** post-election audit sample size methods. Stark simply uses a value of "1" or 100% for his MLU, but then negates this ultraconservative (and unreasonable) assumption of $MLU=1$. (Study Stark's methods more closely to see this.) Saltman's method uses an MLU of " $0 < s < 1$ " or " $2s$ " and an upper margin error bound of " v ", the number of votes.

REVIEWER #3: "...for which she supplies no clear rationale and which seems unlikely to appeal to advocates of either approach."

ME: If this reviewer's statement were true then ***all*** authors of PEA sample size methods (with the exception of Stark) are likewise lacking in clear rationale because *all other* authors multiply their assumed MLU (Saltman's initial approach) times their expression for the maximum within audit unit margin error. Look more closely at the Aslam et al. methods for example. The rationale for doing this was amply given by Saltman, and by numerous other authors including myself, and repeated in my article in the MLU section on page 7. Unfortunately Verified Voting, the ASA, Common Cause, VoteTrustUSA, the Brennan Center, the LWV, US, etc. are still recommending using the imprecise and sometimes impossible $\$v\$$ expression as the measure for maximum margin error, along with a maximum level of undetectability of $\$2s\$$ thus recommending insufficient sample sizes that do not perform as claimed.

Here is exactly the rationale for using an MLU from my submitted article:

"A smart perpetrator would miscount at most some maximum rate k of the margin overall or would cause some maximum rate k where $k : 0 < k < 1$ of margin error within any one audit unit because if all available votes were switched to count for the perpetrator's candidate then all voters who had voted for another candidate would immediately know that the election results were incorrect.

Thus we assume a maximum level of undetectability k , a maximum rate of margin error, such that if more than k times the upper margin error bound occurs, it would look suspicious and cause immediate action by election officials or by candidates and their supporters.[38, appendix B]

A risk-limiting audit design assumes that a maximum rate k of the upper bound for possible margin error within audit units or, for individual ballot"

REVIEWER #3: "Saltman's approach assumes that if v votes (ballots) are cast in a precinct, some proportion s of those votes might be miscounted without being detected, so the maximum error bound (net change) in each precinct is $2sv$."

ME: First you defend Stark's use of the actual upper margin error bounds that Stark copied from myself and/or Aslam et al. Yet now you defend Saltman's original method that was a major advance for his time and an excellent first step in the discovery process but is now known to be mathematically less accurate, producing insufficient PEA sample sizes. The Saltman's " $2v$ " was first replaced and improved by myself with " $2b$ " in 2006 and now is replaced and improved by the use of accurate upper margin error bounds shown in my article.

Thank you for prompting me to add a section that more explicitly explains the flaws in Saltman's original use of the v or $2v$ approach for estimating maximum margin error. (I am very respectful and appreciative to Saltman whose method was a huge advance at the time, but improvements are needed.) Calculations should use the *actual* amount of margin error possible within each audit unit.

REVIEWER #3: "Stark's approach -- which Dopp calls the "upper margin error bound" -- assumes that every vote in a ..."

ME: It is not appropriate to call upper margin error bounds that I originally recommended in December 2007 before him "Stark's approach" especially because Stark sometimes negates the use of these error bounds in his papers with his methods and mixes up the error bounds of various candidate pairs. Stark should take more care to cite other authors' work that precedes his own to make fewer incorrect characterizations against other authors' works.

REVIEWER #3: "Precinct may have been cast for any of the apparent losers, so the error bound with respect to any particular winner-loser pair is $v + w - l, \dots$ "

ME: The upper margin error bound for a particular winning-losing candidate pair (if we want to detect erroneous over and under-votes) is *not* $v + w - l$. It is " $b + w - l$ " (using the number of ballots cast, b , rather than the number of votes cast, v) We do not want to allow perpetrators an avenue to subvert audits by using under and overvote rather than switching votes between the winning and losing candidates.

The number of ballots cast, not the number of votes cast determines the true actual accurate upper bound of margin error bound.

REVIEWER #3: "...where w and l are the votes initially recorded for that winner and loser respectively. Dopp in effect combines these bounds:..."

ME: Huh!*

The maximum level of undetectability first mentioned by Saltman is *not* an upper margin error bound. Saltman correctly used a maximum level of undetectability times an imprecise and sometimes impossible estimate for the proportion of the total available margin error that we assume the perpetrator is willing to cause. Saltman's margin error bound recommendation is not precisely correct, hence its replacement is now recommended with the upper margin error bounds.

I never combined Saltman's bound of " $2v$ " with "Stark's method". My methods have always been more precise than either Stark or Saltman's methods, and what you call "Stark's error bounds" were first recommended by myself in December 2007, months before Stark adopted them after being informed about my work and I assume after reading my paper that recommended using them.

I substituted an accurate precise expression for the upper margin error bound, " $b + w - l$ ", for Saltman's less accurate approximation of the margin error bound " $2v$ " or " v " (Note that the "2" is an attempt to inaccurately translate the number of votes to a measure for margin error), but I rederived and always used Saltman's original

idea of a maximum level of undetectability, which I may have been the first person to independently re-discover in May 2006.

REVIEWER #3: "Thus we assume a maximum level of undetectability k , a maximum rate of margin error, such that if more than k times the upper margin error bound occurs, it would look suspicious and cause immediate action by election officials or by candidates and their supporters." (p. 4)

ME: Yes true.

REVIEWER #3: "Dopp credits Saltman for the MLU idea, yet elsewhere criticizes $2sv$ as likely to yield inadequate sample sizes."

ME: It frequently occurs that researchers in the process of developing a new field, get a partly correct solution that needs improvements in part. Hence it is necessary to discard the incorrect parts and to use the correct parts of other researchers' recommendations, just like with the PPMEBWR and PPMEB methods originally suggested by Calandrino et. Al and Aslam et al., as well as the methods suggested originally by Saltman, several components of which needed improvements.

I independently rediscovered Saltman's idea of a maximum level of undetectability in late May or early June 2006, with help from an appendix in a May 2006 Brennan Center report, but immediately made a small but important improvement, by using the number of ballots cast, rather than votes counted. Stark's unreasonable assumption that the maximum level of undetectability should be "1" or "100%" of course is discarded as well as Saltman's use of the inaccurate upper margin error bound of " $2v$ " because it does not take into consideration the precise within audit unit vote shares that determine the actual margin error that is possible.

REVIEWER #3: "Although Dopp says that she uses an error bound based on $v + w - 1$ (for instance, see the..."

ME: Not true. I would never recommend using the expression " $v + w - 1$ " for the upper margin error bound because it ignores the possibility of improperly counted under and over-votes, thus leaving the loophole for perpetrators that other authors leave open by recommending using the quantity " v " rather than " b ". In my article, I clearly give the upper margin error bounds as " $b + w - 1$ " for any specific winning-losing candidate pair.

REVIEWER #3: "...denominator of the first equation for M/E on page 7), in effect her error bound is $k(v + w - 1)$, where k is a constant analogous to $2s$."

ME: Yes. True. Good observation.

REVIEWER #3: "I see no rationale for this hybridization."

ME: It is difficult for me to imagine why anyone could "see no rationale" for using precise, rather than inaccurate within audit unit upper margin error bounds rather than the less precise and sometimes impossible estimate of " $2v$ " that was first proposed in 1978 that does not account for under and over-votes, and does not make use of the detailed within audit unit vote shares for candidates that are necessary for calculating the correct margin error bounds.

I simply improved upon the method that I derived back at the end of May 2006 that was similar to, but a small improvement to, Saltman’s original method.

REVIEWER #3: “Saltman’s approach can be justified by assuming that we have prior expectations about the relative performance of the candidates in various precincts.”

ME: Not true. No one working in this field, including Saltman, has used “expectations relative to performance” to develop or to justify any of PEA randomly selected sample size methods thus far. We now know exactly how much margin error at most could occur within each reported audit unit by looking at the initial detailed audit unit reported vote counts and ballots cast. No assumptions about prior expectations are used to calculate the randomly selected sample sizes.

REVIEWER #3: “If a candidate’s vote share in a precinct is, say, 20 points higher than we would expect based on covariates, then we have reason to be suspicious of the result. (This argument raises many important questions: Do we even have access to relevant covariates? How reliable are they? Assuming that our “suspicion algorithm” is not failsafe, how much risk do we accept by relying on it? And how can we ensure that “suspicion” leads to correction?)”

ME: PEA random sample size calculation methods do *not* use any “suspicion algorithms” to make random selections. I.e. none of the sample size recommendations by any authors claim to calculate a ‘suspicion’ level for each audit unit.

Suspicion level (exit poll-like calculations against some expected measure) come into play during PEAs only to select the discretionary (non-random) audit units that candidates add to the randomly determined sample.

Most authors’ methods for determining the random sample currently try to determine the maximum possible error that reasonably could occur “without raising suspicion” to determine what sample size is necessary to be able to detect miscounted audit units. Methods that assume a maximum level of undetectability that is less than “1” must also rely on candidates selecting (not randomly) additional audit units based on suspicion.

Thus the algorithm that this Reviewer has in mind to determine suspicious looking audit units is a necessary, but as yet un-researched and inadequately discussed aspect of helping the candidates to determine which discretionary audit units to select to have audited in addition to the random sample that is thoroughly discussed in my article.

This selection of discretionary precincts is a good area for future research using the exit-poll-like analysis methods that you mention because it is a key requirement for post-election audit samples to also include any suspicious-looking precincts as well as the randomly selected sample.

REVIEWER #3: “In Dopp’s approach, the MLU is a function of reported vote share regardless of any prior expectations.”

ME: Not true. The “MLU” is *not* a function of reported vote share in any researcher’s work. The MLU is an arbitrarily selected ratio or percentage (at least for now it is arbitrary, later on data may become available to help better estimate

the MLU) that is greater than zero and less than one. This assumption of an MLU makes it logically necessary to allow for discretionary suspicious audit units selected by losing candidates being audited in addition to the calculated random sample.

On the other hand, the upper margin error bounds *are* a function of the reported within audit unit vote shares and that is why the original Saltman method of using “2v” to measure margin error is inaccurate.

REVIEWER #3: In effect, when Dopp uses $k = 0.5$, she assumes that the MLU is half the apparent winner’s initial vote share,

ME: Not true. I have emphatically never recommended using half of the “winner’s initial vote share” for any calculation, nor have I ever recommending multiplying the MLU, $k = 0.5$, times “the winner’s initial vote share” of any candidate as a method. Such an approach would make no sense whatsoever. I initially (years ago) recommended multiplying the MLU times two times the number of ballots cast, but my article that the Reviewer read recommends multiplying the MLU times the actual within audit unit upper margin error bound for a specific winning-losing candidate pair (when calculating sample sizes) or times the within audit unit upper margin error bounds for *all* winning-losing candidate pairs (when calculating sampling weights).

REVIEWER #3: “or in other words, that an apparent winner in a two-person race can double his or her vote share in each precinct without being detected.”

ME: That’s crazy. I never made or inferred in any way such an obviously stupid statement. Perhaps the Reviewer is confusing my recommendation for using correct upper margin error bounds with the idea of comparing expected with initial reported vote shares in order to determine “suspicious-looking” audit units. This is a topic that does need further development in order to help candidates select additional discretionary audit units for auditing in addition to the randomly selected ones, as mentioned in my article in Section “Other Sample Size Considerations – Losing Candidates Select Additional Audit Units”, but not yet discussed in detail in any post-election auditing papers.

Again, all post-election auditing random sample size calculation methods try to determine how much margin error could at most exist in each audit unit – and *not* how suspicious-looking each audit unit is.

REVIEWER #3: “Why? Why is a change in vote share from 10% to 30% more likely to arouse suspicion than a change from 50% to 100%?”

ME: This a straw-man argument. Obviously I would never make such a ridiculous claim. The topic of how to judge when audit unit results look suspicious has very little to do with calculating a random sample size. The topic of trying to determine when reported audit units look suspicious is very appropriate however in discussing methods to help candidates choose discretionary audit units to add to the audit sample in addition to the randomly sampled audit units. Or perhaps this Reviewer understood “candidate vote share” when I wrote “margin error bound”? I don’t understand the exact source of confusion.

REVIEWER #3: “There may be a reason, but I do not see it, and Dopp does not offer one.”

ME: Not true. I would never try to justify such a ridiculous claim.

REVIEWER #3: “(Dopp does say that k must be < 1 “because if all available votes were switched to count for the perpetrator’s candidate then all voters who had voted for another candidate would immediately know that the election results were incorrect.” That obviously will not suffice as a rationale for using $k = 0.5$.)”

ME: True. However this argument *is* a valid basis for not using $k = 1$, although Stark negates this ultra-conservatism with his arbitrary tolerance level or by calculations that would cause the MLU to cancel from top and bottom of a fraction or from both sides of an inequality.

As stated in my paper the rationale for using $k = 0.5$ rather than the value 0.4 that is currently recommended by Aslam et al. and other authors, is to be more conservative (produce larger sample sizes) than those suggestion using the quantity “ $2s$ ” where “ $s=0.2$ ” because $k = 0.5$ is more conservative due not only to the value of “ k ” itself being greater than 0.40 but also to the fact that “ $2v$ ” grossly underestimates the upper margin error bound in cases where the just-winning, just-losing candidate pair is under consideration, thus under-estimating the sample sizes, or over-estimating the stated probability of detecting error.

REVIEWER #3: “Estimating C

In uniform sampling methods, Dopp proposes to use the average audit unit size to estimate C, the minimum number of miscounted units that could alter the outcome. She recognizes that this assumption “frequently causes C to be overestimated,” and proposes to “adjust for this effect” by using a slightly conservative sample size formula and using $k \geq 0.5$ (page 9). Dopp offers no reason to expect these adjustments to work.”

ME: Not true. I never proposed using “the average audit unit size to estimate C, the minimum number of miscounted units that could alter the outcome”. You incorrectly claim that I am proposing using the old approach of judging the possibility for margin error by the number of total votes counted or cast in each audit unit.

The truth is that I propose using the mean *within audit unit upper margin error bound* to estimate C, the minimum number of miscounted units that could alter the outcome – a method that is significantly more accurate than the method that you incorrectly attribute to me.

My article explains the situations in which such estimates are necessary (whenever data, expertise, and tools are unavailable for more precise estimates) and that larger values of “ k ” produce larger sample sizes.

This new estimate method is a more precise and easy estimate than previous estimates that have been recommended by other authors. Election officials need such easy methods of estimations that they can do themselves for planning purposes, as mentioned in my paper.

REVIEWER #3: “Consider her Example 1 on page 10, treating the 2002 U.S. Senate race in South Dakota. The apparent margin of victory M is just 524 votes. I do not have access to precinct-level counts for this race, but I did find precinct-level results for Minnehaha County (the largest county in

South Dakota) in 2006, with a similar number of total votes cast. In 2006, one precinct had over 1700 votes cast, and others were almost as large. It therefore appears (depending on the actual vote counts and the error bound assumptions) that a miscount in a single precinct could account for the entire margin of victory. Yet Dopp estimates that $1 / C \approx 0.219$, i.e., $C \approx 4.6$. The difference between $C = 1$ and $C = 4.6$ strikes me as consequential, especially when Dopp elsewhere suggests that “risk-limiting audits could eliminate the need for automatic recounts because all sufficiently close-margin election contest[s] would automatically use a 100% sample size whenever necessary...” (page 3).”

ME: Thank you. This comment prompted me to derive another simple formula for the ratio that bounds the minimum number of audit units that it would take to cause an incorrect outcome from my original estimate and thus provide an improved estimate for planning purposes that I’ve added to my article.

I am delighted to report that this derivation took me only about 30 minutes and that the new formula nicely reduces to a very simple formula and improves my estimation method.

However, regarding the example that you gave:

- 1. You do not supply any actual vote counts for that largest precinct, so we can only estimate whether or not that one precinct has the potential to contribute a margin error of 524 votes without raising suspicion or not. I.e. To reverse a margin of 724 votes, 262 +1 votes would need to be switched from the real winner to the apparent winner or some other arithmetically equivalent combination of error would have to occur.**
- 2. If we assume that the margin in that largest precinct is similar to the overall margin and is virtually 50/50 in this one precinct with 1700 votes cast and that there were no undervotes or votes for other candidates in this precinct, then only half or 850 votes out of 1700 are available to be switched to the wrong candidate, and so the Reviewer assumes that almost all of them (263 votes out of 850 or 31% of available target votes that could cause an incorrect outcome that could be switched without raising suspicion in order for this one precinct to contribute all of the margin error.**

Hence, unlike Stark, you are assuming a maximum level of undetectability (MLU or k in my article) of at least $k=0.32$ (but probably much more since the assumption of no under and over-votes and votes for other candidates may be unrealistic -- although you provide no justification for your assumption ;-).

Note, that if there is enough margin error in a single audit unit to cause suspicion then the just-losing candidate should select that audit unit for inclusion as one of his discretionary audit units to be manually counted along with the randomly sampled audit units.

Thanks to this helpful comment.

ME: REVIEWER #3: “PPEB weights

Whereas Dopp has previously argued that sample sizes should be based on the vote counts of the just-winning and just-losing candidates, in her discussion of “weighted sampling methods” she argues that basing sample weights upon a single winning-losing pair is unreasonable and may even be subject to gaming.”

ME: Sort of. Precisely the audit sample size calculations should use correct upper margin error bounds that depend on the number of ballots cast and the vote counts of candidates.

Considering the just-winning/just-losing candidate pair produces the most conservative (largest) sample size but does not produce the most conservative sampling weights.

If this same error bound were incorrectly used for sampling weights then audit under-scrutinizing the just-losing candidate' votes but scrutinizing the votes of the winning opponent and other losers more. Any audit using such sampling weights might be subjected to court challenge and perpetrator strategies.

REVIEWER #3: "Dopp explains, "if a perpetrator knows that selections are weighted to scrutinize margin error that hurts or helps particular candidates, then the perpetrator can try to avoid detection of vote fraud by making a real winning candidate look like a different initial losing candidate and a real losing candidate look like a different initial winning candidate than our weights are focused on" (page 10). It is not immediately obvious how such an attack would work,..."

ME: Thank you for taking so much time to explain in detail your understanding.

Let me try again to communicate with an example. Eg. In a multi-winner contests incorrect sampling weights could possibly allow a perpetrator to make the rigged winner the winner with the most votes, rather than the just-winning candidate in order to avoid scrutiny of audit units containing illicit votes for the illicit winner, or in a single or multi-seat contest a perpetrator could potentially pile up tons of illicit votes from a real winner-reported-loser to a different real loser such that the real loser would become the just-losing candidate (and the real winner-reported-loser would look like a lesser initial loser -- the 2nd runner-up) such that the votes of the real loser to whom all the illicit votes had been transferred would not be scrutinized, etc. It is particularly important to use correct sampling weights in the PPMEB method for sampling individual ballots.

I find it easier to figure things out than to describe them to others or to come up with examples because I work from my forte in math logic.

If the upper margin error bounds for a *particular* winning-losing candidate pair are used to weight selections then the selections are weighted to ignore votes of the loser in that pair and to weight selections to most closely examine the particular winner of that pair. This approach provides options for increasing the chance of escaping detection of fraud by hiding illicit miscounted votes in the ballots that are weighted not to receive much scrutiny – i.e. To hide votes in the just-losing candidate column and in any winning-candidate column who is not the just-winning candidate.

I hope that this is clearer. Thanks for prompting me to re-explain.

REVIEWER #3: "...but the possibility provides one rationale for the error bound in Stark's CAST paper, which uses the ..."

ME: Stark was most assuredly *not* the first person to derive or recommend using the just-winning/just-losing candidate pair error bounds. Stark was the second person to recommend using them and at least the fifth person to "discover" or discuss them, months after other authors. You should review post-election auditing literature or simply my bibliography for my article before making such claims.

I also initially erred in recommending using the same just-winning/just-losing candidate pair upper margin error bounds for both sample size calculations and for sampling weights. The article that you read corrects that mistake.

REVIEWER #3: "...largest error bound for any winning-losing pair in each precinct. However, Dopp proposes instead to use an error bound (equation (2) on page 7, equation (6) on page 12) that equals $2\sum_i w_i + \sum_i l_i$, where l_i comprises "the number of total votes for any losing candidates plus the number of total under or over-votes" in each audit unit (page 7). (Dopp then multiplies this error bound by her MLU ratio k .)"

ME: Yes. True.

REVIEWER #3: "If this equation means what it seems to mean, it is wildly conservative."

ME: Thank you. I'll take that as a compliment. As it turns out, more accurate methods are more conservative.

REVIEWER #3: "For instance, if 100 ballots are cast in a precinct, and all four candidates -- two apparent winners and two apparent losers -- initially receive 50 votes apiece, Dopp's error bound (before applying k) is 300 votes."

ME: I think that you mean that 200 total ballots are cast, and Yes, I agree that $2(100)+100 = 300$ votes is the accurate upper margin error bound.

REVIEWER #3: "This error bound can be interpreted as a comparison of the initial outcome "50 votes for one candidate, 150 votes for the rest" to the hypothetical alternative "200 votes for one candidate, 0 for the rest" -- which, of course, is impossible."

ME: Very Good. Your specific example makes it very easy to disprove your claim of "impossibility" by showing how a 300 vote margin error can occur in your own example. Below I provide one out of the 12 possible examples for how a 300 vote margin error that contributes to altering election outcomes (who the winners are) can occur in your own example.

Assume that candidates A & B are the initially, but incorrectly reported winners, and that candidates C & D are the initially incorrectly reported losers. Given this scenario there are four possible scenarios that result in a 300 vote margin error.

Miscount the initial votes as follows in this precinct.

- 1. The 50 votes for candidate A should really have been counted for candidate C, causing a 100 vote margin error for the pair A & C.**
- 2. The 50 votes for candidate D should really have gone to candidate C, causing a 50 vote margin error for the pair A & C.**
- 3. The 50 votes counted for candidate B should really have been counted for candidate D, causing a 100 vote margin error for the pair B & D.**
- 4. The 50 votes counted for candidate C should really have been counted for candidate D, causing a 50 vote margin error for candidate pair B & D.**

Add it up for a 300 vote margin error in this scenario.

For each of the 6 possible pairs of winners and losers possible for this reviewer's scenario, there are 2 such examples of a 300 vote margin error, for 12 total examples

to refute the claim that it is “impossible” for there to be a 300 vote margin error in this scenario.

These examples are simple to generate using tables listing all the upper margin error bounds for each possible winning-losing candidate combination and selecting the combinations that can occur simultaneously and produce a 300 vote margin error.

No higher margin error can be produced because my formula gives the exact upper margin error bound for all winning-losing candidate pairs.

REVIEWER #3: “Or perhaps Dopp means something different. In any case, this entire issue of weights needs closer explication.”

ME: You understood exactly the precise formula that I proposed.

ME: CONCLUSION

Computer scientists especially understand the concept of “Garbage in – garbage out” --- that it is not possible to produce good outputs by using incorrect inputs.

There is wide-spread confusion and misinformation on how to calculate PEA sample sizes among alleged experts in post-election auditing.

I appreciate the Reviewers’ taking their time to present their points and give me an opportunity to improve my paper and to more clearly correct some widespread misunderstandings about post-election audit sample size methodology.

An unfortunate consequence of the widespread confusion on post-election auditing methodology has been that influential well-intentioned groups such as Common Cause, Citizens for Election Integrity, Minnesota, the League of Women Voters US, Verified Voting, VoteTrustUSA, The American Statistical Association, The Brennan Center, The Florida Voters’ Coalition and other similar groups have been promulgating recommendations for post-election audits that fail to achieve their stated probabilities for detecting incorrect election outcomes and consequently many recent state statutes and procedures have fallen short of the mark.

It is important that improved effective post-election auditing methods are promulgated to replace the older less precise methods in this evolving field.

I do very much appreciate the Reviewers for taking their time to comment so that I can improve my own communications and rewrite my post-election auditing papers to more clearly explain the how methods for calculating post-election audit sample sizes can be improved, the procedures that must be followed, the flaws in today’s voting system designs with respect to auditability, and how to analyze discrepancies found in an audit in order to determine whether to certify an election contest or to expand the audit.

I thank the Reviewers very much for their many helpful detailed comments.