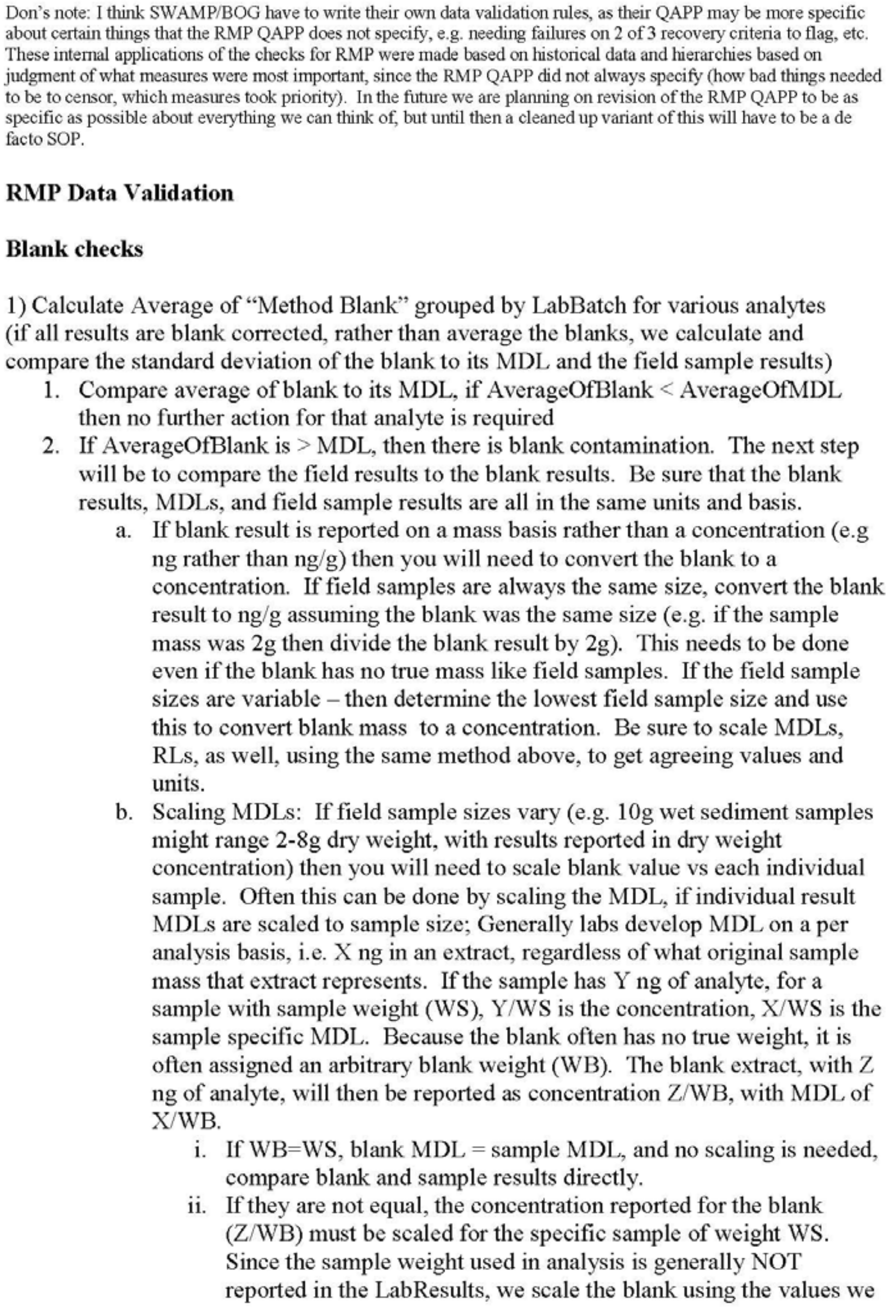
Appendix VI: SFEI Procedures

SFEI Procedures			
Page	Procedure/Equipment	SOP Number	Revision Date
A	RMP Data Validation		

Appendix VI A: RMP Data Validation



Don's note: I think SWAMP/BOG have to write their own data validation rules, as their QAPP may be more specific about certain things that the RMP QAPP does not specify, e.g. needing failures on 2 of 3 recovery criteria to flag, etc. These internal applications of the checks for RMP were made based on historical data and hierarchies based on judgment of what measures were most important, since the RMP QAPP did not always specify (how bad things needed to be to censor, which measures took priority). In the future we are planning on revision of the RMP QAPP to be as specific as possible about everything we can think of, but until then a cleaned up variant of this will have to be a de facto SOP.

RMP Data Validation

Blank checks

1) Calculate Average of "Method Blank" grouped by LabBatch for various analytes (if all results are blank corrected, rather than average the blanks, we calculate and compare the standard deviation of the blank to its MDL and the field sample results)

1. Compare average of blank to its MDL, if AverageOfBlank < AverageOfMDL then no further action for that analyte is required
2. If AverageOfBlank is > MDL, then there is blank contamination. The next step will be to compare the field results to the blank results. Be sure that the blank results, MDLs, and field sample results are all in the same units and basis.
    a. If blank result is reported on a mass basis rather than a concentration (e.g ng rather than ng/g) then you will need to convert the blank to a concentration. If field samples are always the same size, convert the blank result to ng/g assuming the blank was the same size (e.g. if the sample mass was 2g then divide the blank result by 2g). This needs to be done even if the blank has no true mass like field samples. If the field sample sizes are variable – then determine the lowest field sample size and use this to convert blank mass to a concentration. Be sure to scale MDLs, RLs, as well, using the same method above, to get agreeing values and units.
    b. Scaling MDLs: If field sample sizes vary (e.g. 10g wet sediment samples might range 2-8g dry weight, with results reported in dry weight concentration) then you will need to scale blank value vs each individual sample. Often this can be done by scaling the MDL, if individual result MDLs are scaled to sample size; Generally labs develop MDL on a per analysis basis, i.e. X ng in an extract, regardless of what original sample mass that extract represents. If the sample has Y ng of analyte, for a sample with sample weight (WS), Y/WS is the concentration, X/WS is the sample specific MDL. Because the blank often has no true weight, it is often assigned an arbitrary blank weight (WB). The blank extract, with Z ng of analyte, will then be reported as concentration Z/WB, with MDL of X/WB.
        i. If WB=WS, blank MDL = sample MDL, and no scaling is needed, compare blank and sample results directly.
        ii. If they are not equal, the concentration reported for the blank (Z/WB) must be scaled for the specific sample of weight WS. Since the sample weight used in analysis is generally NOT reported in the LabResults, we scale the blank using the values we

DO have reported: sample concentration (Y/WS), sample MDL (X/WS), blank concentration (Z/WB) and blank MDL (X/WB). What we want to derive is Z/WS, the concentration that the blank would have been if it were exactly the same weight as the sample. We can do that from the 4 reported values with some algebra, let us know if you need the equation written out.

c. If the field result (in ng/g) <3*AverageOfBlank (scaled) then flag field sample with VRIP, (censored result- blank is likely too large a component of field result to be quantitative), superceding any existing IP. (We discussed in yesterday's phone call using 5*AverageOfBlank – this would be more conservative and would result in more rejected data – SWAMP should make the decision on using 3* or 5*).
d. If field result (in ng/g) >3*AverageOfBlank flag field sample with VIP if not already IP flagged.

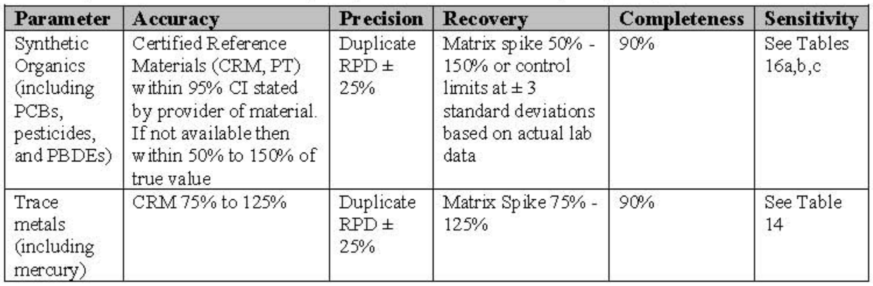
Accuracy check

Keep in mind that if the BOG QAPP specifies other range and/or failure requirements (e.g. BOTH CRMs and MSs must fail to get results flagged/censored), BOG should follow its stated requirements.

In RMP we use a hierarchy for accuracy checks,

SRM > MS > BlankSp

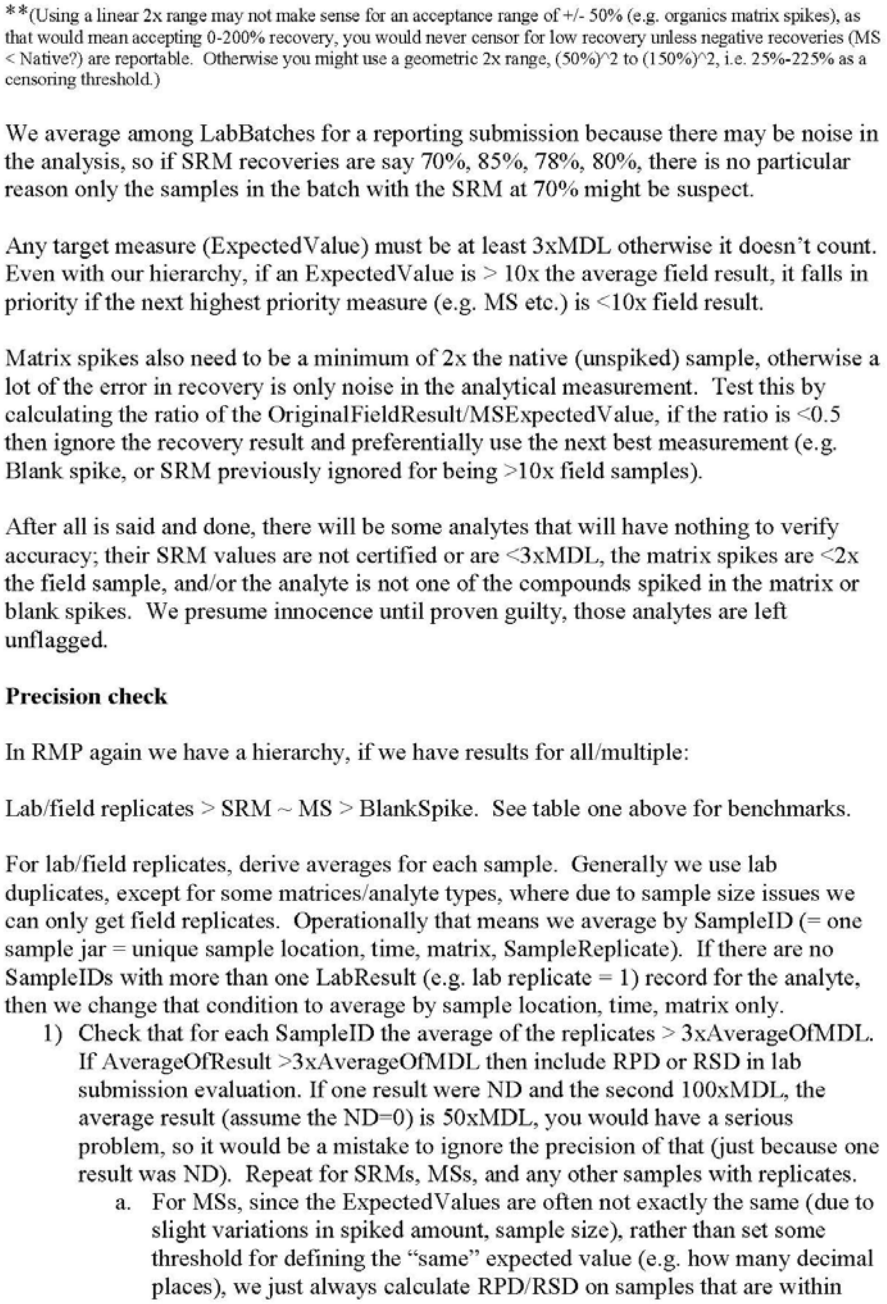
Table 1. (Table 11 from BOG QAPP) shows BOG Data Quality Objectives



Parameter	Accuracy	Precision	Recovery	Completeness	Sensitivity
Synthetic Organics (including PCBs, pesticides, and PBDEs)	Certified Reference Materials (CRM, PT) within 95% CI stated by provider of material. If not available then within 50% to 150% of true value	Duplicate RPD ± 25%	Matrix spike 50% - 150% or control limits at ± 3 standard deviations based on actual lab data	90%	See Tables 16a,b,c
Trace metals (including mercury)	CRM 75% to 125%	Duplicate RPD ± 25%	Matrix Spike 75% - 125%	90%	See Table 14

If there is a higher priority measure (e.g. SRM) within an acceptable range then the analyte passes or fails regardless of the outcome of the remaining measures (e.g. MS).

We average across LabBatches for a project submission (e.g. one year of data for one lab considered together). Moderate failure (> target range but <2x the range – see table 1 above) of the highest priority usable measure get a VIU assigned to each analyte that fails the test, a bad failure (>2x outside target range) gets VRIU, which gets applied to the field samples, indicating that this reporting batch (which may be several LabBatches) has suspect values for the given analyte.**



**(Using a linear 2x range may not make sense for an acceptance range of +/- 50% (e.g. organics matrix spikes), as that would mean accepting 0-200% recovery, you would never censor for low recovery unless negative recoveries (MS < Native?) are reportable. Otherwise you might use a geometric 2x range, (50%)^2 to (150%)^2, i.e. 25%-225% as a censoring threshold.)

We average among LabBatches for a reporting submission because there may be noise in the analysis, so if SRM recoveries are say 70%, 85%, 78%, 80%, there is no particular reason only the samples in the batch with the SRM at 70% might be suspect.

Any target measure (ExpectedValue) must be at least 3xMDL otherwise it doesn't count. Even with our hierarchy, if an ExpectedValue is > 10x the average field result, it falls in priority if the next highest priority measure (e.g. MS etc.) is <10x field result.

Matrix spikes also need to be a minimum of 2x the native (unspiked) sample, otherwise a lot of the error in recovery is only noise in the analytical measurement. Test this by calculating the ratio of the OriginalFieldResult/MSExpectedValue, if the ratio is <0.5 then ignore the recovery result and preferentially use the next best measurement (e.g. Blank spike, or SRM previously ignored for being >10x field samples).

After all is said and done, there will be some analytes that will have nothing to verify accuracy; their SRM values are not certified or are <3xMDL, the matrix spikes are <2x the field sample, and/or the analyte is not one of the compounds spiked in the matrix or blank spikes. We presume innocence until proven guilty, those analytes are left unflagged.

Precision check

In RMP again we have a hierarchy, if we have results for all/multiple:

Lab/field replicates > SRM ~ MS > BlankSpike. See table one above for benchmarks.

For lab/field replicates, derive averages for each sample. Generally we use lab duplicates, except for some matrices/analyte types, where due to sample size issues we can only get field replicates. Operationally that means we average by SampleID (= one sample jar = unique sample location, time, matrix, SampleReplicate). If there are no SampleIDs with more than one LabResult (e.g. lab replicate = 1) record for the analyte, then we change that condition to average by sample location, time, matrix only.

1) Check that for each SampleID the average of the replicates > 3xAverageOfMDL. If AverageOfResult >3xAverageOfMDL then include RPD or RSD in lab submission evaluation. If one result were ND and the second 100xMDL, the average result (assume the ND=0) is 50xMDL, you would have a serious problem, so it would be a mistake to ignore the precision of that (just because one result was ND). Repeat for SRMs, MSs, and any other samples with replicates.
    a. For MSs, since the ExpectedValues are often not exactly the same (due to slight variations in spiked amount, sample size), rather than set some threshold for defining the "same" expected value (e.g. how many decimal places), we just always calculate RPD/RSD on samples that are within

usable range (if MS ExpectedValue > 3xMDL, and Native/MS ExpectedValue < 0.5)***

2) Average across LabBatches for a project submission (e.g. one year of data considered together) for each SampleType that has replicates. Results that average <3xMDL for an analyte are ignored

3) Use the hierarchy for replicates to choose the SampleType for assigning QAcodes If there is a higher priority measure (e.g. lab/field duplicate) within an acceptable range then the analyte passes or fails regardless of the outcome of the remaining measures (e.g. SRM or MS). Similar to accuracy, we flag on a project/event/submission basis for one lab, as there will be some variation in average concentration and noise around the RPDs/RSDs- a higher RPD in one batch may just be a result of the sample chosen for that rep having lower concentration, or just the odds of heterogeneity in subsamples of a grab sample, etc. If on average there is often a problem with precision for an analyte, then we have a problem that should be noted/flagged.

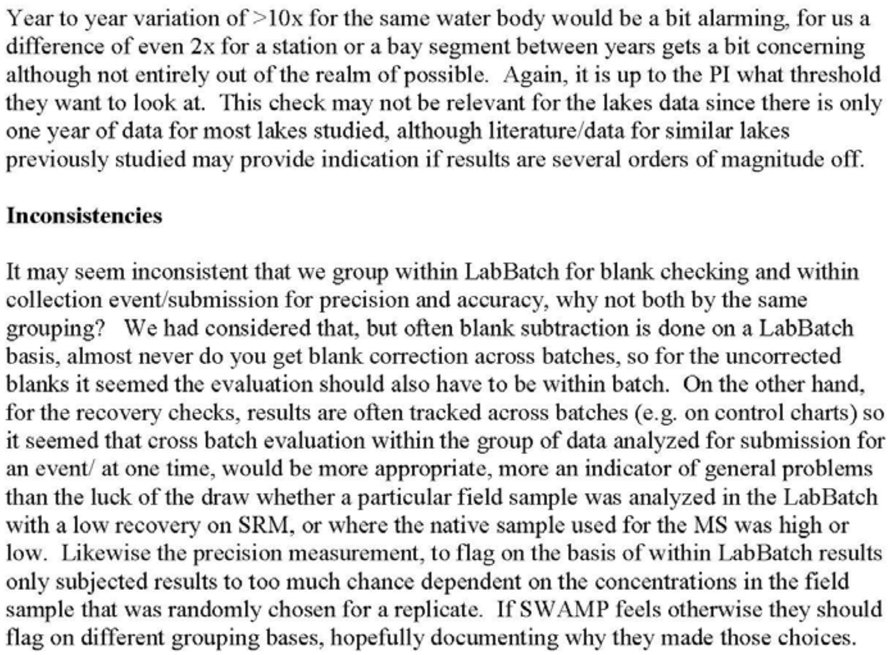
4) If RPD or RSD for an analyte (averaged across batches for a lab submission, for the SampleType chosen for evaluation) is >25% AND <50% (>1x to <2x target range) apply VIL to that analyte for the entire submission. If RPD or RSD >50% (>2x range) then apply VRIL to that analyte for the entire lab submission.

*** SWAMP outlines a method for doing an MS/MSD RPD check by different methods depending on whether the spiking value is the same or different- if spiking value is different, RPD is calculated on %recovery, if spiking value is the same, RPD is calculated on the raw values. For RMP we just go for calculating on %recovery always (worst case scenario, avoids some problems of trying to judge when to go for one calculation method versus the other)- e.g. you have a parent sample with 5 ng/g, and you make an MS/MSD spiking 10ng each sample, which turn out to 0.9 and 1.0g- are the expected values the same or not? Depends on how much you round – one you expect a concentration of 5ng/g*1g = 5ng +10ng = 15ng in 1g (expectedvalue=15ng/g.) The other you get 5ng/g*0.9g = 4.5ng +10ng = 14.5ng in 0.9g = 16.1ng/g. What if the MS vs MSD is 0.95 vs 1.0 g mass? There will often be some inexactness in the MS/MSD masses and spiking volumes, always going by RPD on %recovery avoids those questions. Otherwise where is the cutoff where the ExpectedValue is considered equal? Within 1%? 2%? 5%?

Range check

This may not be possible with the lakes data as they are distinct water bodies, so expected concentrations my not fall within some predictable range like in SF Bay where conditions are similar and waters mix.

Nonetheless, differences in concentration of >100x within a species for similar water bodies might be something of a concern. What the review threshold should be is a bit of professional judgement. We generally contact the reporting lab to make sure if they stand by the number. Sometimes they fix the number (seeing matrix interference, ion ratio issues, peak drift, etc on closer examination of the raw data), other times they say it should remain as is. Then it is a matter of professional judgment whether we believe it, if not, we flag VQ or VRQ depending on whether we think the number is possible/realistic, or so far out of range that it should be censored.



Year to year variation of >10x for the same water body would be a bit alarming, for us a difference of even 2x for a station or a bay segment between years gets a bit concerning although not entirely out of the realm of possible. Again, it is up to the PI what threshold they want to look at. This check may not be relevant for the lakes data since there is only one year of data for most lakes studied, although literature/data for similar lakes previously studied may provide indication if results are several orders of magnitude off.

Inconsistencies

It may seem inconsistent that we group within LabBatch for blank checking and within collection event/submission for precision and accuracy, why not both by the same grouping? We had considered that, but often blank subtraction is done on a LabBatch basis, almost never do you get blank correction across batches, so for the uncorrected blanks it seemed the evaluation should also have to be within batch. On the other hand, for the recovery checks, results are often tracked across batches (e.g. on control charts) so it seemed that cross batch evaluation within the group of data analyzed for submission for an event/ at one time, would be more appropriate, more an indicator of general problems than the luck of the draw whether a particular field sample was analyzed in the LabBatch with a low recovery on SRM, or where the native sample used for the MS was high or low. Likewise the precision measurement, to flag on the basis of within LabBatch results only subjected results to too much chance dependent on the concentrations in the field sample that was randomly chosen for a replicate. If SWAMP feels otherwise they should flag on different grouping bases, hopefully documenting why they made those choices.