Author’s response to reviews

Title: Wrangling Environmental Exposure Data: Guidance for Getting the Best Information from your Laboratory Measurements

Authors:

Julia Udesky (udesky@silentspring.org)
Robin Dodson (dodson@silentspring.org)
Laura Perovich (lperovich@gmail.com)
Ruthann Rudel (rudel@silentspring.org)

Version: 2 Date: 22 Oct 2019

Author’s response to reviews:

We thank both reviewers for this second reading of our manuscript and the additional helpful comments.

Below, please find our point-by-point response to the reviewer comments.

Reviewer #1:

Comment 1.1: The manuscript is much improved. It is coherent and insightful and highly readable. I have a few minor quibbles with the text on quantitation methods that should be addressed.

Response: We appreciate the positive comment about our improved manuscript and the assistance correcting the section on quantitation methods.

Comment 1.2: On line 184, the external method of quantitation is listed as simple and flexible. I don't understand the use of the term "flexible". I would never use that term, as there is no flexibility in the need to have a calibration curve that spans the expected concentration range of samples- whether the analyses are being done by an external or internal standard calibration method. Rather say that the method is straight-forward.

Response: Thank you for pointing out that “flexible” was not the right word – we meant to say that external calibration can be used in a variety of analyses because it does not incorporate some of the requirements of the other methods discussed (e.g. does not require that the standard be a compound not present in the samples). We revised the sentence to read:
Lines 180-183: External calibration, where the response (i.e., chromatogram peak) from the sample is compared to the response from calibration standards containing known amounts of the analytes of interest, is a simple method that can be used for a variety of different analyses.

Comment 1.3: In discussing the external method the authors mention that it does not correct for instrument performance. That sentence is misleading as it implies that the problem is with the instrument. The instrument carries out the same actions... but it cannot control for matrix effects that govern the transfer efficiency of analytes from the injector to the column. This is a sample extract problem, not an instrument problem. There can be a considerable difference in transfer efficiency and/or degradation between pristine standards and environmental samples in the injector and there is no way to detect this when using external calibration.

Response: We corrected our description to clarify that the variation in instrument response is a result of the matrix effects:

Lines 183-185: “However results can be influenced by interference from other chemicals present in the sample matrix and resulting fluctuations in the analytical instrument response.”

Comment 1.4: In addition(line 193), I am not familiar with researchers using an internal standard with the isotope dilution method. The method (software) simply corrects for all analytical losses by having the isotope that is added at the start of the extraction procedure function as both an SRS and an IS.

Response: It is interesting to hear that this reviewer is not familiar with this practice. We have heard from other chemists about this practice and also found it described in EPA Method 8000D (https://www.epa.gov/sites/production/files/2015-12/documents/8000d.pdf, p. 32). We made some minor edits, shown below, to try to clarify.

Lines 190-195: “Finally, for isotope dilution methods — which are the most accurate — labeled isotopes for each of the target compounds are added to samples prior to extraction. An additional internal standard is added to the samples just prior to injection to monitor loss of the labeled isotopes in the extraction, and the analytical software then corrects for loss during sample extraction and for effects of the sample matrix (e.g., presence of other compounds in the sample that interfere with the analysis).”

Reviewer #2:

Comment 2.1: Thanks to the authors for the changes that much improve the flow and comprehensibility of the manuscript. There is such useful material here that it serves the reader well to be able to move more cleanly from one section to the next. I greatly appreciate the authors' efforts in creating this summary of methods of QA/QC review as I can see it be immensely helpful to many researchers new to working with outside laboratories. The large quantity of new material and re-organization, however brought up some additional issues for me as listed below. If these are addressed, I see this manuscript as being acceptable for publication
though still tricky to navigate at times. Numbers refer to the track changes version of the document.

Response: Thank you for the additional careful review and helpful suggestions!

Comment 2.2: Please review the document for the use of the term "qualify" as it is still used in a rather technical sense at times. For the more novice reader, it may not be understood as a way of looking at or flagging values between the LOQ and LOD and many may not be used to seeing values provided for this situation. For example, lines 243-5 could read as "flagging" instead of "qualifying". There are quite a few mentions of this before the scenario is explored in 301-315. Line 320 is a little hard to follow (a detect vs. a true detect).

Response: Thank you for pointing out the need to clarify our use of the term “qualify.” We reviewed each instance of this term in the manuscript and, when appropriate, replaced it with the term “flag” (e.g., lines 256, 267). We also:

• Made more explicit what we mean by “qualify” at the first mention of the term (Lines 64-68):
“Despite these helpful documents, there is still a lack of readily accessible, practical guidance on how to interpret and use the results of both field and laboratory quality control (QC) data to qualify exposure datasets (i.e., flag results for certain compounds and/or certain samples as imprecise, estimated, or potentially over-or-under-reported) and this gap is reflected in the environmental health literature.”

• Revised the sentence on Line 320 to read:
“If the lab reports data qualifier flags, it may also be necessary to clarify the interpretation of those flags, including but not limited to which flags distinguish non-detects from detect results above the MRL and estimated values.”

Comment 2.3: In general, please check over the document for possible confusions between laboratory measurement problems and overall method measurement problems. The authors refer to this in line 234 but do not explicitly address the difference.

Response: We kept this comment in mind while reviewing the manuscript and made the following changes:

Lines 301-304: Replaced “revealed data concerns” with “revealed systematic problems with the analytical method.
“For newly developed methods where performance is not characterized, we request results for all recoveries of spiked samples and/or CRMs so that we can perform visual checks that have at times revealed systematic problems with the analytical method that were not noted by the lab (see Data Interpretation: “Is the method accurate?” for discussion).”

Lines 489-491: Specified the source of bias in numeric values could be from contamination in the field or lab or from bias in the analytical method:
“After we establish the MRL for chemicals that are detected in blanks, we are confident that levels in samples above that value are true detects and that they are correctly ranked, but there may still be concern about consistent bias in the actual numeric values being reported, both from contamination in the field or lab or from bias in the analytical method.”

Comment 2.4: Authors have reorganized and added information to present the batching issues related to QA/QC review more cleanly. However do they state that researchers should work with labs to develop a batching plan? They also don't address whether they would request a lab to review and compare different batch results before releasing data back to the researchers.

Response: We have not often worked with lab analysts to develop a batching plan. Rather, by specifying the order in which samples should be analyzed, we try to separate field QC samples enough that they end up distributed across batches, while still maintaining lab blinding to the field QC. We agree that it could be useful to discuss batching with the lab but do not have particular practical advice to offer based on our own experience. We added a note suggesting that it could be helpful to discuss batch size and composition with the lab:

Lines 392-401: “Two approaches help address batch-to-batch variability: (1) randomizing participant samples between batches by specifying the order and grouping of samples (and blind field QC samples) when submitting samples to the lab (this may require corresponding with the lab to determine the batch size in advance), …”

We also typically do not request that the lab review/compare different batch results before releasing the data. Given that some batch issues are revealed by field QC samples (to which the lab has hopefully remained blinded), we have found it reasonable to first consider batch issues ourselves and then follow up with the lab with questions as needed.

Comment 2.5: Line 288: I think it's more correct to say the MRL is lowest level that some labs will report or report without flagging/qualifying as estimated. Again, since a lab may not be providing values below the LOQ.

Response: We made this revision:

Lines 255-257: “The reporting limit (RL) or method reporting limit (MRL), which is either the lowest value that the lab will report or the lowest value that the lab will report without flagging the data as estimated, is often (but not always) the same as the quantitation limit or LOQ.”

Comment 2.6: Line 309: Does the term "non-detect" need to been defined (as meaning below the LOD)? Remember that some groups may only have received an MRL cutoff.

Response: We revised the sentence to define non-detect as below the detection limit.

Comment 2.7: Lines 366+: What LCS do the authors’ suggest for blood? Bovine serum brings a different suite of measurement issues, including potential background levels of some analytes of interest.
Response: We clarified with the following changes:

Lines 307-312: “If CRMs aren’t available, the laboratory can prepare laboratory control samples (LCS) by spiking known amounts of target chemicals into a clean sample of the matrix of interest, such as a dust wipe, air sampler, purified water or synthetic urine or blood that has been analyzed and shown to be free of the analytes of interest, or to contain a consistent amount of analytes of interest that can be subtracted from the amounts measured in the spiked sample to calculate a percent recovery.”

Comment 2.8: Line 76: I think the authors mean measure invalidity; not that the measurements are invalid. The tests may be fine, but the conclusions drawn are incorrect.

Response: We revised the sentence as advised:

Lines 68-70: “While the vast majority of environmental health studies report robust findings based on high quality measurements, questions about measure validity have led to confusion and lack of confidence in some topic areas.”

Comment 2.9: Line 180: Matrix choice does not fit cleanly into a "chemical identity" section. Matrix choice is also an issue beyond biological matrices and plays a role in environmental sampling as well.

Response: We moved the matrix choice to its own paragraph with the sub-heading “Matrix” and added a comment that matrix choice is relevant for environmental sampling.

Lines 159-164: “Matrix: Another consideration in developing the analyte list is what type of samples are available (if working with stored samples) or will be collected. As discussed previously, certain biological matrices are preferred over others for measurement of certain chemicals (e.g., reference 9). Matrix type is also relevant for environmental samples; for example, physical-chemical properties like the octanol air partitioning coefficient inform whether an analyte of interest is more likely to be found in air or dust (Dodson et al. 2014).”

Comment 2.10: Line 183: Is this paragraph meant to fall under the "chemical identity" sub-section? Might it more be "method performance"?

Response: We agree that this paragraph didn’t fit under the “chemical identity” sub-section. We combined two sections (“What can we measure?” and “How will the analysis be performed?”) into one (“What can we measure and how?”) and gave this particular paragraph its own heading, “Method” (lines 165-1772)

Comment 2.11: Lines 196-7: The method of choice for the researcher is based on the levels of concern to the researcher, not necessarily that they can enumerate all the anticipated concentrations.

Response: We agree with this point and have revised the sentence on lines 202-204 to read:
“Another important factor in selecting a method is to make sure the method is sensitive enough to detect the anticipated concentrations in the field samples down to levels that are relevant to the research question.”

Comment 2.12: Line 340: The 20% recommendation is not terribly clear - 20 field QC vs 80 field samples, 20 field QC samples vs 100 field samples, 20 field QC vs 80 other QC?

Response: Thank you for pointing out the need for clarification here! We revised the recommendation as follows (lines 285-288):

“We base the number of QC samples we collect in the field on budget and our sample size, generally aiming for at least 20% QC samples (e.g., if collecting 80 field samples then collect 16 field QC samples), though a higher percentage are needed in small studies.”

Comment 2.13: Lines 362 and 370: Please keep introduced terminology consistent: include "standard". Line 370 should be LCS not "spiked sample". Likewise, lines 366-369 could use the term "spiked" so as to clue the reader to the adherence to the paragraph header. Please check that paragraph beginning at 571 also has consistent terminology.

Response: Thank you for the careful read and for noting these inconsistencies! We made the requested changes and also corrected an error in the way we referred to reference material – in most cases we should have referred to certified reference material (CRM) rather than standard reference material (SRM), as SRM is a special kind of CRM produced by NIST (https://www.nist.gov/srm/srm-definitions). The revised paragraphs are copied below:

Lines 303-316: “Spiked samples and certified reference material: Spiked samples and certified reference material (CRM) establish the accuracy of the method by assessing the recoveries of known amounts of each target chemical. A CRM is a matrix comparable to that used for sampling (e.g., drinking water) that has been certified to contain a specific amount of analyte with a well-characterized uncertainty. If CRMs aren’t available, the laboratory can prepare laboratory control samples (LCS) by spiking known amounts of target chemicals into a clean sample of the matrix of interest, such as a dust wipe, air sampler, purified water or synthetic urine that has been analyzed and shown to be free of the analytes of interest. The LCS or CRM – at least 1 per analytic batch – are run through the same sample preparation, extraction, and analysis as the field samples to capture the accuracy of the complete method; calculating the percent of the known/spiked amount recovered for each analyte tells us whether the method is accurate in the matrix.”

Lines 450-459: “Spiked samples and certified reference material: Box 1 outlines our approach for analyzing spike recovery or CRM data. The approach is similar for all of these samples. However one distinction is that if LCS recovery and other QC measures, such as lab blanks (matrix, solvent method, or other) are acceptable, a poor matrix spike recovery (higher or lower than acceptable bounds) can alert chemists to interferences from matrix effects, and suggest steps to address this such as matrix-matched calibration (17). We typically only use data for analytes that have average LCS, matrix spike or CRM recoveries between 50%-150%, though this decision criterion can be adjusted based on the needs of the project. If we do retain data for
chemicals with spike or CRM recoveries outside of the acceptable range, we note in publications that concentrations in our data may be under- or over-reported.”

Comment 2.14: Lines 396-406: It is possible that readers may be more confused by the difference between surrogate recovery standards and internal standards (since both are added to every sample) than with matrix spikes (which could be more easily contrasted by starting each section saying one addresses overall recovery and the other individual sample recovery).

Response: We added a comment about how surrogate recovery standards differ from internal standards and revised the first few sentences to clarify the contrast between surrogate recovery standards vs. spiked samples and certified reference material.

Lines 340-348: “Surrogate recovery standards: Whereas recoveries from LCSs, matrix spikes and/or CRMs tell us about the performance of the method in a clean or representative matrix, surrogate compounds are used to evaluate recoveries from individual samples. Recoveries of surrogate compounds can help identify any individual samples that may have inaccurate quantification, for example due to extraction errors or chemical interferences. Surrogates, like internal standards, are spiked into each sample, however surrogates are added prior to sample extraction to assess the efficiency of this process. Internal standards, on the other hand, are added after extraction, just prior to injection into the chromatographic system, to account for matrix effects and other variation in the instrument response during analysis.”

Comment 2.15: Lines 428-431: This addition about blood sample field blanks is difficult to follow. I assume the authors mean to state what methods can be used in lieu of blood sample blanks, though they may be inadequate.

Response: We revised the discussion of blood and urine sample field blanks as follows:

Lines 371-379: “Unfortunately, in some cases there aren’t good options for representative field blanks. For example field blanks can be created for biomonitoring programs by taking empty collection containers into the field and using purified water or synthetic urine or blood to create a blank (4). However, important short-comings of this approach are that (1) it is difficult to capture contamination that can be introduced by sample collection materials such as needles and plastic tubing used to collect blood, (2) water may not perform the same as urine or blood in the extraction and analysis, and (3) the lab will likely be able to identify the field blanks. Similarly, it is difficult to maintain lab blinding when using a “clean” matrix like vacuumed quartz sand as a field blank for vacuumed house dust.”