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: 1 Date: 19 Aug 2019

Author’s response to reviews:

We thank both reviewers for their careful reading and constructive comments. We have incorporated many suggestions from both reviewers and hope that our commentary is now better organized and more accessible for our intended audience. In addition to the revisions requested by the reviewers we also (1) checked the consistency of our use of the terms “QC” and “QA/QC” (2) made minor revisions (in tracked changes) to the abstract and main text for clarity (3) made minor revisions (in tracked changes) to a few of the guideline “boxes” (Box 0, Box 1 and Box 4) and (4) added another example table to the supplementary information (Additional file 2).

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

Reviewer #1:

Comment 1.1: The authors have taken on the very challenging objective of trying to capture the critical elements of QAQC that need to be included in an exposure study involving both field and laboratory measurements. As they have pointed out in their Background and in Additional File #1, inadequately performed or reported QAQC data limit the ability of the scientific community to objectively assess the quality of measurements purported to link environmental exposures with health risks. In the current world of ever-tighter funds for research, we need to be vigilant so that data reported are adequately supported by QAQC analyses, and adequately interpreted in light of the QAQC results, to obviate the need for repeating studies. The manuscript is based on a series of questions that a researcher may reasonably ask when trying to plan a study and analyze data, so this feels like an interesting conversational approach. Very useful and practical topics are covered in an easily readable format. And the use of concrete examples from their own studies helps to make essential points in a clear manner.
Response: We thank the reviewer for the positive comments!

Comment 1.2: I do find, though, that sections tend to wander over topics not specifically called out in the header and thus it comes across as unfocussed. Sometimes information is presented in several sections that really should be included under one heading. For example, in the section "Before collecting data: what can we measure?" there is a listing of the different terms that are used to define detection and quantification limits (DL, MDL, LOD, LOQ, PQL, LQL, CRQL), and a paragraph about use of the RL, RL/2 and 0 as substitutions for the term non-detect. Then under the section "After collecting data: what was measured?" there is a discussion of the use of ND (not detected) and "0" in reporting data. But the use of these two terms is clearly linked to detection and quantification limits, and so it seems that there needs to be a tighter section that deals with all of these terms and their appropriate use and the reporting of which term(s) was/were assessed.

The key issue of S:N needs to be mentioned and discussed because this is fundamental to analytical measurements. In addition, researchers get sloppy at times and do not distinguish between instrument detection limits and method detection limits, and which is being reported. These are different concepts and need to be clarified for the expected audience.

Response: We appreciate the general comment about organizational issues in this paper. We made considerable revisions to the organization by:

- Breaking our guide up into three distinct phases (Study Design, Study Implementation and Data Analysis) and providing a brief road map early on to indicate how the paper is organized (lines 116-119):

  “Our guide is organized by a series of questions that we ask when we start a new study and then again when we receive measurement data from the lab. Key QA/QC concepts are introduced in the Study Design section and are most thoroughly addressed in sections about Study Implementation and Data Interpretation.”

- Adding sub-headings throughout the paper to help readers easily locate key QC concepts that appear in more than one section.
We also appreciate the advice about how to improve our discussion of detection limits. We moved the discussion of different terminology to a new section, “How will the lab report the data?” and added a discussion of the distinction between instrument and method detection limits with reference to signal-to-noise ratio (lines 233-240):

“Common terms used by laboratories to discuss reporting limits include instrument detection limit (IDL), method detection limit (MDL) and limit of quantitation (LOQ). The IDL and MDL are both related to the level of an analyte that can be detected with confidence that it is truly present. The IDL captures the smallest true signal (change in instrument response when an analyte is present) that can be distinguished from background noise (variation in the instrument response to blank samples), while the MDL takes into account additional sources of error introduced during sample preparation (e.g., the extraction process, possible concentration or dilution of samples) and thus is higher than the IDL.”

Comment 1.3: As an aside, the sentence beginning on line 160 of the former section does not make sense. The instrument detection limit is a defacto limit, and there should not be any sample with a detection or quantification below this limit, and thus no need to “censor” such imaginary data. Please understand and use analytical terminology correctly.

Response: Thank you for pointing out our improper use of “instrument detection limit.” We meant to refer to the lab’s reporting limit. We removed the reference to the IDL in this paragraph, which now reads (lines 169-176):

“An important factor in selecting a method is to make sure the method is sensitive enough to detect the anticipated concentrations in the field samples. For example, commercial labs measuring environmental chemicals may establish reporting limits to meet the needs of occupational or regulatory safety compliance testing; these limits may be much higher than levels that are meaningful for research questions about general population exposure and could result in most data being reported as non-detect or qualified as estimated and imprecise. On the other hand, lower reporting limits generally translate to more expensive testing, so researchers have the opportunity to balance sensitivity and cost.”

Comment 1.4: Other times a section presents information that should be included under an entirely different rubric. Under the section "Before collecting data: what can we measure?" the
discussion starts with identifying analytes, getting detection limits, identifying QAQC samples, and how reports will be formatted. Of those 4, only the first topic seems to fit under the rubric of "what can we measure". The others are discussions of "how will the analyses will be conducted": targeted or non-targetted identification; "what type of quantification will be used": external standard, internal standard, or isotope dilution; "what QAQC samples and procedures are needed": blanks, spikes, duplicates, surrogates, SRMs; and "what data will the lab provide to me". The normalization of data as discussed at the end of the first paragraph belongs under "what data will the lab provide to me". The discussion of isotope dilution analyses for biological matrices is misplaced under the "what QAQC to plan for". This is a type of quantification, not QAQC. The topic of "what can we measure" is critical to more fully address with respect to multi-residue/multi-analyte analysis programs. It is a whole different kettle of fish to analyze for 1-2 compounds of similar structure vs 100 compounds of differing compound classes, polarities, solubilities, and volatilities. This is an excellent opportunity to draw on text scattered throughout the manuscript that addresses this issue.

Response: We appreciate the feedback about the “What can we measure?” section. We made the following revisions:

• We removed the discussion of targeted vs. non-targeted identification. The second reviewer found our discussion to be too high level to be useful to the intended audience, and we think a more substantial discussion is beyond the scope of this commentary. Instead, we acknowledge in an earlier section that non-targeted methods exist, but that our commentary is primarily relevant to data from targeted analyses (lines 127-130).

• We retained the discussion about establishing appropriate reporting limits, as we consider ‘what’ we can measure to include both the types of chemical and the levels that can be measured.

• We moved the sentence about normalization of data to the “How will the lab report the data?” section.

• We moved the discussion of isotope dilution analysis to the new section “How will the analysis be performed?” section and added a brief description of quantitation with external calibration, internal calibration and trade-offs.
The second paragraph of the “What can we measure?” section now highlights the challenges associated with developing a multi-residue analytic method (lines 156-163):

“The process of determining a final list of analytes will differ depending on whether the lab has an established method or is developing a new method, and whether it is targeted to a few chemicals with similar structure versus many chemicals with different properties (different polarities, solubilities, etc.). Targeting a broad suite of chemicals may limit the degree of precision and accuracy that can be achieved for each individual chemical, and the lab may need to invest substantial effort to develop a multi-residue method – that is, a method that can analyze for many chemicals at once – and determine a final list of target chemicals with acceptable method performance.”

Comment 1.5: Be careful in the use of "name" when describing a chemical (e.g., line 234). Need to distinguish between chemical names and trade names because errors can creep in when trying to analyze for compounds that only exist in the industrial sector where rigor may not be enforced and misunderstandings can ensue.

Response: We appreciate this point and added a recommendation to avoid the use of trade names (lines 150-153):

“One approach is to send the lab a list of the chemical names (avoiding the use of trade names, which can be imprecise), Chemical Abstracts Service (CAS) numbers, and configurations (e.g., branched or linear, if relevant) of all desired analytes (see additional file 1 for example correspondence).”

Comment 1.6: As an aside, the sentence beginning on line 139 is incorrect and misleading. An internal standard is used to enhance quantification accuracy, not to "confirm that a particular chemical or chemicals of interest are present".

Response: Thank you for the correction! We moved this sentence to an earlier section and revised it as follows (lines 125-127):
“Our guidance is most relevant to targeted organic chemical analyses, which use liquid or gas chromatography, often in combination with mass spectrometry, to determine whether a predefined set of chemicals are present in samples.”

Comment 1.7: The issue of limiting contamination, which is described in the final paragraph of "what can we measure" is misplaced in this section. The issue of assessing contamination through the use of blanks could be an important section on its own. This encompasses both lab and field QC. But it is not the duty of the analytical lab to figure out how to address collection and shipment contamination, as implied in the sentence starting on line 181. There is a need to discuss selection of materials for shipping and storage... very practical issues such as selecting pure PE containers or using muffled aluminum foil. These tips to new researchers would be useful and helpful. And please avoid the use of the "plastic", as it is jargon for polymeric material.

Response: Thank you for the helpful suggestion about the placement of the discussion of preventing contamination. We moved this discussion into its own section, “How to minimize sample contamination?” and expanded it by adding specific examples (those noted by the reviewer and some additional examples from CDC guidance) (lines 204-224).

We elected to keep the word “plastic” as we think it is more understandable/accessible than “polymeric material” for the intended audience.

Comment 1.8: The terms for specific QAQC sample types needs to be corrected, as some used are not sufficiently specific (e.g., "lab blank"). For example, a solvent blank is just the extraction solvent concentrated from initial volume used for extraction down to the final analysis volume. The solvent method blank is the extraction solvent carried through the entire process, including reagents and cleanup sorbents and cartridges and such. The matrix blank is the sampling matrix (gauze, PUF) carried through the method. And by analogy, there are solvent method spikes and matrix spikes... the terms "lab control samples" of "lab spikes" are too ambiguous. Similar terms for field QAQC samples need to be checked for clarity.

Response: We agree that the term “lab blank” is not helpful for referring to a particular type of blank (a solvent method blank, solvent blank, matrix blank, etc.). We tried to only use the term “lab blanks” when we were referring to the combination of blanks provided by a lab (which
could include one or several types of blanks) and to distinguish them from field blanks. Our remaining references to “lab blanks” are as follows:

Lines 507-512: “Unexpected findings, such as a chemical or chemicals detected in much higher levels in a lab blank (matrix, solvent method, or other) than in the field blanks, warrant further investigation. In this case, we might suspect that the lab blank was contaminated by another sample; examining the sample run order (which must be requested from the lab, see example correspondence in additional file 2) could shed light on whether a very high sample was run directly before the lab blank.”

Lines 429-432: “The distinction is that if the lab spike 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).”

Lines 480-482: “We use detected levels in field and lab (matrix, solvent method, or other) blanks to evaluate whether to raise the method reporting limit, qualifying some of the values reported by the lab that are in the range of the blanks.”

We removed all references to “lab spikes” but retained the references to “lab control samples”. What we are referring to as a lab control sample – spiking into a clean sample matrix known to be free of target analytes – could be referred to as a “matrix spike,” but we’re concerned that we are then lacking a distinct term for the type of spike that is created by spiking into a representative field sample. We found the term “lab control sample” in this guidance document (http://www.publications.usace.army.mil/Portals/76/Publications/EngineerManuals/EM_200-1-10.pdf?ver=2013-09-04-070852-230) and this EPA method document (https://www.epa.gov/sites/production/files/2015-12/documents/8000d.pdf) and we have also experienced labs using this term when reporting their QC data. We agree that this term is ambiguous but we ultimately could not determine a better way to distinguish between these different types of matrix spike samples.

Comment 1.9: The discussion of surrogate recovery standards (SRSs), which runs throughout the manuscript needs to be strengthened and corrected. Per line 342-343, an SRS is not an internal standard. Likewise, it is incorrect to say "should avoid using chemicals from different classes as
The writer does not appear to understand the function, selection and use of internal standards as separate from surrogate recovery standards. The sentence starting at line 352 is also incorrect. Isotope dilution software corrects for losses, but when using SRSs it does not (typically). Again, a correct discussion of quantification methods and the trade-offs is warranted. Analyte and SRS recoveries > 120% should be discussed as a special issue.

Response: We corrected the reference to an SRS as an internal standard and the statement that isotope dilution software corrects for losses when using SRS. We appreciate any additional feedback from further review of our commentary about whether we have succeeded in providing a correct discussion of SRSs and how analyte and SRS recoveries > 120% should be discussed.

Comment 1.10: Other egregious errors that must be corrected are related to use of the term noise. From line 421… "we're reporting are signal and not noise from measurement error". Noise is a very specific issue in electronic signals and it is not related to measurement error. From line 452… "data is explained by "noise" or imprecision". Again, this is not a correct use of the term noise.

Response: We removed all instances where we used the term “noise” to describe measurement error.

Comment 1.11: The concept of this manuscript is good, but the organization needs to be significantly modified for coherence. The technical terms need to be used and described correctly and with specificity. Publish with major revisions.

Response: We believe the re-organization and correction of technical terms as described above have improved the quality of our manuscript.
Comment 2.1: This manuscript contains valuable insights into the intermediary steps between lab analysis of environmental samples and epidemiologists or exposure scientists use of results in hypothesis driven statistics. It also provides a very useful set of steps for systematically evaluating the quality of data received and what interpretations or assessments of confidence can be made. Many points described are not performed systematically across the field and will be useful additions to scientist work plans.

My main concern for this document is the rather confusing organization of material, which includes the same concepts/ideas being discussed in multiple locations throughout the text, as well as Boxes and Additional Files. While it looks like the authors are trying to match the text flow to their internal data processing protocols, I found this organization confusing for the reader. It's of course very tricky because there is a lot of interplay between the topics. I might suggest more focus on each QAQC type in turn: defining, describing how to plan for, and providing instructions for how to use each element all in the same section, but this would be a big reorganization. (I found making an outline this way helpful for my understanding of your arguments.) An example is having QAQC definitions provided in sections that are 'after receiving your data' -- seemed counterintuitive. In addition, I think the scope may be too wide, but I leave that to the authors.

Response: Thank you for the very helpful feedback about the organization of the paper. We made considerable revisions to the organization by:

• Breaking our guide up into three distinct phases (Study Design, Implementation and Analysis)
• Providing a brief road map early on to indicate how information is organized in the paper (lines 116-119):

“Our guide is organized by a series of questions that we ask when we start a new study and then again when we receive measurement data from the lab. Key QC concepts are introduced in the Study Design section and are most thoroughly addressed in sections about Study Implementation and Data Interpretation.”

• Adding sub-headings throughout the paper to help readers easily locate key QC concepts that appear in more than one section.
• Moving QAQC definitions to the “Study Implementation” section.

Comment 2.2: The manuscript struggles a bit with having a clearly defined audience. I'd suggest focusing on early career exposure scientist/env epis and the scientists from other disciplines who are dipping a toe into exposure science, and not try to include citizen scientists and more senior level scientists who haven't been using all their QAQC data to its potential. Perhaps because of this wide aim of audience, there are some abrupt shifts in tone between introductory level material (definitions of LOD and LOQ) and more technical aspects (qualification of estimated values below the RL.)

Response: We appreciate this comment and agree that early career exposure scientists and environmental epidemiologists who do not have training in analytical chemistry are our primary audience. We added the following to try to help clarify our audience (lines 133-135):

“While it may be most useful to environmental health scientists who have little or no experience in analytic chemistry, we hope that researchers with a range of experience will find it helpful to consult our approach for evaluating and presenting QC data in publications.”

Comment 2.3: There are also sections with very high level discussion (which may be too high level to be of use to the reader, such as the discussion of nontargeted analyses) and others that provide a sudden glimpse of detail in an otherwise overview paragraph (for example lines 116 - 117 seem out of place in an introductory paragraph and could go in the Experimental Design section.) These latter types of items make it difficult to read because the reader doesn't know when an item is going to be fully addressed.

Response: We have worked to make the level of detail more consistent throughout the manuscript. With regard to the specific example cited by the reviewer, we reduced the discussion of non-targeted analyses, with the goal of acknowledging that these methods exist but that our commentary is primarily applicable to data from targeted analyses.
Comment 2.4: I think most of the steps discussed are certainly useful to the reader, but I think it would very useful to also discuss what types of QAQC analyses will already have been done by whichever type of lab the authors are using and which will have not. Some labs do many of the steps the authors discuss and it may be worth mentioning that the authors will be confirming (duplicating) their efforts. It’s also a bit unclear where the authors’ purpose is to alert the reader on how to catch lab mistakes (for example, a method that’s not performing well) and where there are elements beyond the lab's procedures that are the responsibility of the contracting scientist. Perhaps the authors are trying to use less confrontational language, but for the less aware it could be confusing.

Response: We tried to be more explicit about different needs for QA/QC depending on whether the lab is developing a new method or using a well-established method, and to explain why we might duplicate some elements of the laboratory review. Relevant revisions in the “What QAQC is needed?” section include:

• Lines 163-168: “In any case, a new method should be validated to characterize performance measures – precision, accuracy, expected quantitation and method detection limits, and the range of concentrations that can be quantitated with demonstrated precision and accuracy – before analyzing study samples. If the lab already has an established method for the chemicals of interest, the research team should review method performance measures to ensure they are consistent with study objectives.”

• Lines 283-287: “While laboratories generally conduct rigorous review of their own QC data, considering lab and field QC together can help to identify specific sources of contamination, imprecision, and systematic error, so we typically request to review the lab’s raw QC data in conjunction with the field QC data.”

• Lines 316-319: “For newly developed methods where performance is not characterized, we request results for all recoveries of spiked samples and reference materials so that we can perform visual checks that have at times revealed data concerns not noted by the lab (see Data Interpretation: “Is the method accurate?” for discussion).
Comment 2.5: I do notice that most of the examples are about batching issues and I think this is a very underestimated issue in exposure science/environmental epidemiology. I would argue for increasing the prominence of this and grouping some of the batch related discussion. I think the authors could dedicate a little more room to this in the Experimental Design section and move other parts into it (especially lines 267-271). I think it would be helpful to add some discussion on talking with the lab about sizes of analytic batches and requesting particular ordering of samples. Otherwise, apportioning the QAQC elements, as instructed, is a bit mysterious. The authors may want to make it clearer how "batch" is being used throughout the article (some labs have multiple batches within an instrument's run and some may term the whole run a batch). Readers may confuse it with a single set of results they receive from the lab, if they received more than one.

Response: We re-organized the batch discussion, tried to clarify what we mean by the word “batch”, and discussed a few ways that we have approached apportioning QAQC across batches. The batch discussion is now organized as follows:

In Study Implementation (lines 371-387):

“Analytical batches: Analytical performance can shift over time and even between multiple extractions or instrument runs within a short time window. Laboratories often analyze samples in multiple batches and the time between batches can vary from days to months or even years, though ideally this time span is minimized in order to maintain consistent equipment and procedures throughout the study. A “batch” is a set of field samples (samples submitted to the lab) and associated laboratory QC samples that are analyzed together in one analytic run.

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) in each batch ahead of time, and (2) running standard reference materials (SRM) – such as from NIST (26) – in each batch of samples in order to characterize drift. When standard reference materials are not available, another option is for the researcher to prepare identical/split reference samples. We have done this, for example, by pooling together several urine specimens and making many aliquots of the pool, then including 1-2 blinded samples from this pool with each set of samples we send to the lab. If the laboratory analysis is performed in multiple batches, all QC elements should be examined on a batch-specific basis. Not every laboratory will specify whether or not samples were analyzed in batches; it is a good idea to request that a variable for batch be included in the results report.”
In Data Interpretation (lines 409-421):

“Analytical batches: Examining results by batch or even by sample run order can reveal trends in QC samples over time, identifying systematic laboratory errors that may be missed by summary statistics or visualizations (20). Shifts in method performance over time may require batch-specific corrections or dropping or qualifying data from certain batches. Notably, a trend in QC sample results over time can be problematic even if they remain within the acceptable limits established by the lab. In our own work, for example, examining our data by analytic batch revealed an upward trend in sample-specific detection limits for some analytes, such that detection limits in later batches were within the range of sample results from earlier batches (Figure 1). The detection limits in the later batches still met the specifications of our contract with the lab, but it was clear that we would not be able to compare results in the latter two batches to those in the first three. We showed this plot to the lab and they agreed to re-analyze the samples in the later batches, which resulted in consistent detection limits across all the field samples.”

Comment 2.6: Since the authors spend the bulk of their time talking about environmental samples, I might focus the paper on this. Duplicates, field blanks, lab blanks, and method reporting limits have different difficulties for biological samples which the authors do not really cover. This also makes the main example of poorly presented QAQC (solely on biological sample-based studies) seem inconsistent.

Response: We appreciate this suggestion. Ultimately we decided to retain the material related to biological samples and so we added additional information about QAQC with biological samples as shown below:

Study Design - What can we measure:

• Lines 153-155: “For biomonitoring, another important step is to establish the appropriate biological matrix for measurement of the chemicals of interest (9) and to determine if the parent chemical or metabolites will be targeted.”
Study Design - How to minimize sample contamination:

- Lines 209-220: “The CDC’s guidance on sample collection and management identifies some possible sources of contamination when analyzing for common chemicals like plastics chemicals, antimicrobials and preservatives in blood or urine. Key considerations, depending on the particular chemicals being targeted, include selecting appropriate collection containers (e.g. glass containers if analyzing for plastics chemicals), avoiding the use of urine preservatives (e.g., parabens, BPA), and providing adequate instructions to participants collecting their own samples (e.g., avoid using antimicrobial soaps or wipes during collection) (4). As noted previously, contamination can also be minimized in biomonitoring of some chemicals by measuring a metabolite rather than parent chemical, and possibly by measuring a conjugated rather than free form of the metabolite (9). In some cases, the lab may need to pre-screen collection containers or other sampling materials to see if they contain any target chemicals.”

- Lines 355-358: “In the case of blood collection, while it is possible to check collection materials for the target chemicals, if they are present at all it isn’t easy to quantify how much of the chemical in the field sample is originating from the collection materials.”

Study Implementation - Spiked samples and reference material:

- Lines 293-296: If reference material samples aren’t available, the laboratory can prepare laboratory control samples (LCS) by adding 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.

- Lines 301-304: “Another type of spiked sample, called a matrix spike, can be used to check the extraction efficiency for a complex sampling matrix that may interfere with the analysis. These samples are typically included if there is concern about interference from the sampling matrix, for example, with house dust, soil or sediment samples, consumer products, or biological samples like blood.”

Study Implementation - Blanks:
• Lines 351-353: “Unfortunately, in some cases there aren’t good options for representative field blanks. For example using purified water to create a blank for urine can reveal to the lab which sample is the field blank and water may not perform the same as urine in the extraction and analysis.”

Study Implementation - Duplicates:

• Lines 361-365: “Duplicate samples can also be created by collecting a single sample and splitting it prior to analysis, which is the only option for biological samples; however, this method only captures the precision of the analysis process (14, 17) and could lead to un-blinding of the lab analyst, if for example the split samples are noticeably smaller than others.”

Comment 2.7: Line 68 - I would state what the aspect of confusion was

Response: We agree that it is more helpful to be explicit here. We added sentences describing reasons for lack of confidence in some measures (lines 64-72):

“While the vast majority of environmental health studies report robust findings based on high quality measurements, questions about measurement validity have led to confusion and lack of confidence in some topic areas. For example, a number of studies have measured rapidly metabolized chemicals such as phthalates and bisphenol A (BPA) in blood or other non-urine matrices, despite the fact that urine is the preferred matrix for these chemicals. Phthalates and BPA are present at higher levels in urine and, when the proper metabolites are measured, there is less concern about contamination from external sources, including contamination from plastics during specimen collection (9).”

Comment 2.8: Line 36 - I would include "of" - grams of urinary creatinine, grams of serum lipid

Response: We made this correction.

Comment 2.9: Lines 158, 160 - Since information above this was introductory, I would introduce what the authors mean by qualifying and censoring. Some researchers may only get values and
"ND"s, so may not understand that the authors mean an "ND" to be a qualification (if I am interpreting the authors correctly.)

Response: We removed uses of the word “censor” as we agree that we were using it to mean different things (report as non-detect vs. remove data from a dataset) and it was creating confusion. We have been more specific about what we mean by qualifying as in lines 243-244:

• “The reporting limit (RL) or method reporting limit (MRL), the lowest value that the lab will report without qualifying the data as estimated…”

Comment 2.10: Lines 162-165 - The authors use an occ compliance lab's reporting limit as an example of where there would be censoring applied to values below the instrument limit of detection, whereas I believe that is an example of a reporting limit above the instrument's limit of detection.

Response: Thank you for noting our incorrect reference to the instrument detection limit! We meant to refer here to the lab’s reporting limit and to the fact that with a higher reporting limit, more results will be reported as ‘non-detect’ than with a lower limit. We revised the text accordingly so that this section now refers consistently to the “reporting limit”.

Comment 2.11: Line 168 - Linear dynamic range is introduced as a probable lab deliverable but never discussed.

Response: We removed the language about linear dynamic range and stated more explicitly that the goal would be to ask the laboratory for “the range of concentrations that can be quantitated with demonstrated precision and accuracy” (lines 164-165).

Comment 2.12: 194-195 "Provide raw QC data" - couldn't this be terribly extensive. You might want to help people understand how to reign it in.
Response: We tried to be more specific about what types of raw QC data we request from the lab and how the particular data we request varies with the analytical method. Relevant revisions in the “What QAQC is needed?” section include:

- Lines 316-321: “For newly developed methods where performance is not characterized, we request results for all recoveries of spiked samples and reference materials so that we can perform visual checks that have at times revealed data concerns not noted by the lab (see Data Interpretation: “Is the method accurate?” for discussion). For well-established methods, and particularly when isotope dilution quantification is used, it is sufficient to request a table summarizing the spike recovery or reference material recovery results (across batches, if relevant) for reporting in publications. ”

- Lines 334-338: “For analyses using external or internal calibration, we ask the lab to provide us with the recovery results for each surrogate in each sample, so that we can flag any samples or compounds that might have had extraction problems. However if the lab used isotope dilution quantification, we are less concerned about obtaining this raw data from the laboratory given that the reported results are already automatically corrected for extraction and matrix effects.”

Comment 2.13: Lines 200-206 - This is kind of a distracting amount of detail.

Response: We agree that the discussion of isotope dilution was too detailed in the context of the original section in which it appeared. At the suggestion of the first reviewer we moved this discussion to a new section about establishing what type of quantification will be used – we think it is important for the reader to have a general sense of the different types of quantification as this will also influence the type of lab QC that is of interest.

Comment 2.14: Lines 230-241 is a bit repetitive of the beginning section on establishing identifications of analytes. Could be made briefer.

Response: We reduced the length of this discussion and tried to eliminate repetition of the earlier section except where absolutely necessary – we do think it is important to emphasize the particular elements to re-check with the lab at this stage (chemical name, CAS number, isomer).
Comment 2.15: Line 260 - This seems to be the first appearance of "sample-specific detection limits" and was not included in the LOD definition section.

Response: Thanks for noticing this! We now introduce sample-specific detection limits in the “How will the lab report the data?” section (lines 246-250):

“Before submitting samples for analysis, it is helpful to find out (1) the methods and terminology that the laboratory will use to describe reporting limits (LOD, LOQ, etc.) (2) whether reporting limits will be consistent within a chemical or whether limits could vary between samples or batches (3) how the laboratory will report values for compounds with a confirmed identity but measured at levels below what can be accurately quantitated.”

Comment 2.16: Line 287 - I'd suggest spelling it out a little more for the reader - that the huge range in spike recovery evidenced poor accuracy of the method for that chemical.

Response: We revised this sentence as suggested. It now reads:

“However, the range of spike recoveries was ranged from -2% to 1670% and averaged about 750%, indicating that the method was not able to accurately quantify this chemical. Therefore we did not report any data for this compound.”

Comment 2.17: Line 307-311 - Parts of this might be added to the QAQC design section.

Response: We agree that this discussion of approaches to creating matrix spikes belongs in a different section – we moved it to the “What QAQC is needed?” section.

Comment 2.18: Line 356-362 - Are these concerns taking into account the auto-correction or not, it's unclear.
Response: We clarified that our discussion of interpreting surrogate recovery standards is not applicable if the method involved isotope-dilution with auto-correction (lines 451-452).

Comment 2.19: Line 418–419 - This now reads that censoring is not qualifying whereas before it was presented as a type of qualification. (?)

Response: We agree this was confusing as the type of censoring we’re talking about here (removing samples or chemicals from our dataset) was distinct from our earlier use of the word to describe censoring of values below the reporting limit. We removed the word “censor” in both cases and revised this sentence so that it now reads (lines 499-503):

“Once we have decided to keep a particular chemical in our dataset (i.e., if we are confident in the accuracy of the analytical method), we typically qualify values rather than removing measurements from our data. We use field blanks and sometimes other blanks (see Box 3-1) as a basis to qualify data by raising the method reporting limit (MRL), flagging low values as estimated, until we feel confident in the levels we’re reporting.”

Comment 2.20: Line 420 - Is this the first mention of the Method Reporting Limit (other than Table 1)? If so, it should be defined. It could be included in the paragraph on other limit terms.

Response: Thanks for catching this, we added the Method Reporting Limit to the paragraph on reporting limit terminology.

Comment 2.21: Box 0 - there is no reference to this Box in the text?

Response: Thanks for catching this, we added a reference to Box 0 on line 406.

Comment 2.22: Boxes - It's odd to have variable names in these. I presume the boxes are meant to be able to stand on their own?
Response: Thank you for pointing this out! Confirming that currently, these boxes are meant to stand on their own, so we removed the variable names from the boxes.

Comment 2.23: Supp 4 - Since this is labeled a report when the reader prints it out, it reads very strangely to go back and forth between instructions, additional information (to the article reader? to the analyst?), and small blurbs of results (especially when they're in the same font). Perhaps the authors could put boxes around the instruction parts? I fear this Supplement reads currently more as an internal document (pg 12, what is "PrettyName"?) and is not the easiest to navigate as an outsider. It would also be helpful to have consistent colorations in the graphs - sometimes black rims are used on circles and sometimes not, though perhaps I'm missing the point.

Response: Thank you for taking such a careful look at this supplemental document. It has evolved over time and we can see now how the form we submitted was confusing. We significantly revised the document to focus on reporting QAQC results (rather than combining instructions and results), removed references to internal variable names like “PrettyName”. We removed the black outlines from some points and checked for consistency of colors across plots.

Comment 2.24: Table 1 - There are inconsistencies of presentation in this table. For example, for the "Measure" column, I wouldn't use it to supply synonyms or provide instructions. Also, I wouldn't consider "Blank" a measure. Perhaps "quantities detected in blanks" or some such?

Response: Thank you for the careful look at Table 1. We think it is helpful to keep detection limit / quantitation limit synonyms in the table but moved them from the “Measure” column to the “Interpretation” column.