Author’s response to reviews

Title: Fetal growth in environmental epidemiology: mechanisms, limitations, and a review of associations with biomarkers of non-persistent chemical exposures during pregnancy

Authors:

Elizabeth Kamai (ekamai@unc.edu)
Thomas McElrath (tmcelrath@partners.org)
Kelly Ferguson (kelly.ferguson2@nih.gov)

Version: 1 Date: 24 Mar 2019

Author’s response to reviews:

Reviewer #1

Reviewer's Report: The authors present a well-researched and very nicely written summary of a large body of interesting research investigating the influence of non-persistent EDCs on size at birth or fetal growth parameters. Some opportunities to highlight common themes across studies were not fully explored, and the manuscript would benefit from additional work to synthesize the reviewed literature and recommendations for future research directions.

1. There are a few opportunities in this review that I’d like to see the authors explore.

   a. Overall, the topic of possible effect measure modification of the associations by infant sex is a bit diffuse in the manuscript, though the discussion in section 5.3 is robust.

      i. Setting up the need to look at sex-specific associations earlier in the manuscript would improve the flow of the manuscript.

RESPONSE: Thank you for this suggestion. We have added wording in the Potential Etiologic Mechanisms section to introduce the importance of examining sex-specific associations at lines 152-153 (“Furthermore, many of the associations between these compounds and adverse health outcomes demonstrate sex-specific effects”) and lines 177-180 (“Furthermore, the potential involvement of these pathways in the associations between maternal exposure and fetal growth make it imperative that associations in epidemiologic studies be carefully examined for potential effect measure modification by infant sex”). We have also added wording to lines 409-410 to clarify that we were specifically interested in results stratified by sex or restricted to a single sex (“We present results stratified by sex or restricted to a single sex, as the effects of prenatal exposure to some non-persistent environmental chemicals may differ by fetal sex.”).
ii. Additional detail related to observations of sex-specific effects would be beneficial. For example at lines 428-435, clarify whether the associations described among boys versus girls were due to formal assessment of effect measure modification by child sex or due to the study only including male or female infants (like reference 145 was restricted to boys). Studies restricted to one sex could be more clearly designated in the tables.

RESPONSE: We have added additional detail to describe sex-specific effects and evaluation of effect measure modification by infant sex at lines 489-492 (“Of these studies, seven stratified cohorts by infant sex (100-102, 104, 109, 111, 112), five of which reported formal statistical analysis of effect measure modification by infant sex by testing either interaction terms (101, 109, 111, 112) or the difference in coefficient estimates (102). Two cohorts in France were restricted to male infants (97, 98)”), lines 704-712 (“All of these studies reported results of models restricted to male infants; four found statistically significantly positive associations between BPA exposure and birth size in boys”), lines 752-756 (“Interestingly, although three of these studies evaluated effect measure modification by including sex*exposure interaction terms (89, 149, 155), each reported different results of these analyses. One found no significant modification of any effect by infant sex (155), and two found significant modification of the association between 2,5,-DCP or 2,4-DCP and birth weight or length, but in opposite directions by sex (89, 149)”), lines 761-762 (“Three studies included sex*paraben terms to evaluate effect measure modification by infant sex, but none found statistically significant interactions”), lines 764-766 (“A single study found statistically significant effect measure modification by sex of the association between BP3 and birth weight, with BP3 associated with increased birth weight in boys but not in girls (89)”). We have also added shading to the tables to better indicate whether studies restricted to one sex and/or evaluated sex-stratified models.

2. Section 5.1 would be improved by addressing a few additional points related to challenges of exposure assessment of non-persistent chemicals. If you feel strongly that any of these points are outside of your intended scope, then kindly add language to the section that clarifies your intended scope.


RESPONSE: We thank Reviewer 1 for this suggestion and have added a discussion of within-subject pooling of biospecimens at lines 1067-1189 (“Researchers may be reluctant to measure numerous biomarkers during pregnancy due to high cost of laboratory assays. Within-subject pooling of biospecimens, where samples from a single individual at multiple time points are
combined prior to measurement, can be used to address this concern while also reducing misclassification of exposure assessment (203). Increasing the number of biospecimens in an individual’s pooled assay can both decrease bias in the effect estimate and increase power (203). Additionally, within-subject pooling can improve exposure characterization over first morning voids (201). At least 6 and 35 specimens are required to limit bias to 10% attenuation for chemical with ICC of 0.6 and 0.2, respectively, though (203). This number of biospecimens may be unfeasible to collect for logistical or financial reasons. However, if the same number of biospecimens are pooled for each participant, and reliable estimates of ICCs are available, a posteriori disattenuation correction can virtually eliminate bias in effect estimates (203). Moreover, if at least two biospecimens are measured separately, measurement error models such as simulation extrapolation or regression calibration can be used to reduce bias to less than 10% (203). An important limitation to pooling samples across weeks of pregnancy, however, is that key windows of vulnerability to exposure may be missed. Consider a chemical for which exposure during the first trimester is the most relevant for fetal growth and for which there is high variability (low ICC) across pregnancy. If the urine sample from this time point is pooled with those collected later in pregnancy, any potential associations would be diluted.”)

We have included additional recommendations to improve exposure assessment at lines 1206-1209 (“Thus, in an ideal setting, repeated urine samples would be collected within trimesters and across gestation and analyzed individually. Since this is not always feasible financially, alternative approaches—such as exploring windows of vulnerability in a subset and then subsequently pooling—are encouraged.”).

b. Address the importance of considering season of measurement for non-persistent chemicals. My research and that of others has shown evidence suggesting that chemicals like benzophenone-3 and triclosan (and to some degree organophosphate flame retardants) have seasonal exposure patterns. As birth weight also tends to follow a seasonal pattern (admittedly, exactly what this trend is and how strong it is varies given the population under study), it would be worth mentioning that consideration of season of measurement may be important to address in future research. Time of day may likewise be an important consideration (there is some indication of this in Dr. Kate Hoffman’s work on organophosphate pesticides). These factors may obscure or inflate true associations if not considered.


RESPONSE: We thank Reviewer 1 for pointing out this gap in our discussion section and for providing relevant references. We have added a discussion of seasonal variation in exposure and in birth weight, including the references listed, in lines 1444-1449 (“Moreover, there is evidence that exposure to some non-persistent chemicals – such as BP3, TCS, and organophosphate pesticides – follow seasonal patterns (209-212). Birth weight also follows seasonal patterns, though these patterns can vary by population and years under study (213-215). Researchers should critically evaluate (using, for example, directed acyclic graphs (216)), whether season of measurement should be considered in modeling effects of non-persistent chemicals on fetal growth outcomes.”). We have also added a discussion of the variability of urinary measures of non-persistent at different times of day at lines 1050-1055 (“Measuring concentrations of a chemical in a 24 hour urine sample is more representative of the day’s exposure compared to a spot urine sample (204, 205). First morning void samples are more complicated because time of day is a significant predictor of levels of phthalates and BPA in urine, with higher levels of BPA and high molecular weight phthalates observed in samples collected in the evening, and highest levels of MEP in the morning (200, 201, 206).”)

c. Consider specifically commenting on the appropriateness of measuring chemicals in urine collected at delivery (or at least the need to carefully consider incorporating these into exposure averages). My general concern is that the hospital environment may increase exposure to chemicals such as phthalates that may be present in medical tubing used to administer fluids or triclosan/triclocarban which may be present in soaps or other products used frequently in hospital settings. I have not seen much discussion of this in the literature, but I worry that something like high phthalate concentrations measured at delivery may be a marker for a difficult labor and delivery (or proxy for increased medical intervention) which may unduly influence associations observed with birth weight as opposed to measurements taken earlier in pregnancy. It is another important point relevant to temporality issues surrounding exposure assessment raised by the authors.

RESPONSE: We agree with Reviewer 1 that it may be in appropriate to measure exposure to non-persistent chemicals at delivery in an analysis of pregnancy exposure and birth outcomes. We have commented on why such measures should be interpreted with caution in lines 1225-
"Many studies included in this review measured exposure to phthalates, environmental phenols, or other non-persistent consumer products at delivery (107, 109, 111, 112, 154-162, 181, 184, 207), and it was not always clear at what point during delivery urine samples were collected. This timing of exposure should be interpreted with caution. Phthalates are often present in medical devices, intravenous tubing, and medication coating, for example (85). Exposure to these products prior to urine collection could produce higher urinary concentrations of these chemicals or these metabolites, but could not have a causal effect on fetal size at birth. Moreover, the single study that measured phthalate metabolites at delivery (specifically, prior to IV insertion) as well as earlier in gestation reported poor correlation between the two measures for all phthalates, but particularly for DEHP metabolites (109). Even assuming these measures are uncontaminated, they are still may not be representative of earlier, perhaps more relevant, windows of exposure in pregnancy.”).

Minor revisions and clarifications

1. Section 2 on potential etiologic mechanisms should be streamlined.

RESPONSE: We have edited the mechanisms section in order to condense the text and make it more concise and straightforward.

2. Delete "whatever you measured" at line 231.

RESPONSE: This line has been edited accordingly.

3. Was there variability across the studies in terms of the investigators definition of IUGR? This is sometimes defined just a little bit differently by different researchers. Please comment on any discrepancies in definitions across studies or clarify that you excluded studies that did not define IUGR as "estimated fetal weight in the lowest 10th percentile for gestational age" (lines 245-246).

RESPONSE: We agree and have clarified these lines (new lines 404-405, “We included studies that defined IUGR as estimated fetal weight in the lowest 10th percentile for gestational age”), as well as provided additional detail about the two studies that examined IUGR (both described in section 4.1.2. Phthalates and fetal growth outcomes measured during gestation) in lines 580-583 (“Two small hospital-based case-control studies from the same research group reported that levels of DEHP metabolites measured at a single time point in the third trimester were associated with increased odds of IUGR or “fetal growth restriction” (diagnosis of either IUGR or low birth weight) (100, 126).”).
4. Line 279 - although you deal with this in more detail later in the manuscripts, covariates used in multivariable models were also fairly variable across studies, and this would be worth noting/briefly acknowledging both here and in the sections on other EDCs.

RESPONSE: We have added wording about variability in covariates included in models assessed at new lines 457-458 (“covariates included in multivariable models”) and 577-578 (“covariates included in multivariable models,”) for phthalates results, and line 698 (“covariates included in models”) for environmental phenols results.

5. Although they may be familiar terms to most readers, it would be prudent to provide a brief definition of "high molecular weight phthalates" at line 364 and "low molecular weight phthalates" later.

RESPONSE: We have re-written the introduction of the phthalates results section (lines 424-442) to clearly define and describe high molecular weight phthalates and low molecular weight phthalates (“Phthalates are often categorized into two groups based on molecular weight: low molecular weight phthalates are <250 g/mol and include dimethyl phthalate, diethyl phthalate, di-n-butyl phthalate, and diisobutyl phthalate; high molecular weight phthalates include butylbenzyl phthalate, di(2-ethylhexyl) phthalate (DEHP), di-n-octyl phthalate, diisononyl phthalate (89, 90). This classification reflects both structural similarity and similar routes of exposure, as low molecular weight phthalates are often found in personal care and hygiene products, while high molecular weight phthalates are commonly used as plasticizers in polyvinyl chloride (PVC) materials, tubing, medical devices, and food packaging (85, 87).”).

6. At Line 441, specify the specific organophosphate flame retardant that was associated with reduced birth weight in girls.

RESPONSE: We have added the specific chemical name, isopropyl-phenyl phenyl phosphate (ip-PPP) (new lines 793-794).

7. At lines 452-455, similar concerns about at least some non-urinary measures of phthalates are also raised by reference 84. This should be stated in the section on phthalates (or this information should all be saved for section 5.1, but I think the point bears repeating throughout the manuscript as it is an important one).

RESPONSE: We agree and have added sentences at lines 567-571 to indicate why urinary measures of phthalates are preferred (“However, the preferred matrix for measuring human
exposure to phthalates is urine. Phthalate levels measured in other matrices are orders of magnitude lower than levels in urine and more prone to error from contamination (13). Results from these studies are thus not directly comparable to those that used measures phthalate exposure in maternal urine.”).

8. Though I agree with the statements made, you should add some references to lines 622-628. Additionally, it may be of use to cite some studies that have examined intraclass correlation coefficients (ICCs) of the chemicals of interest, particularly where this information is available during pregnancy (I believe reference 91 contains this info for at least some of the compounds of interest).

RESPONSE: We have added a discussion of intraclass correlation coefficients, with appropriate references, to new lines 1037-1042 (“This is exemplified by an extensive literature on intraclass correlation coefficients (ICCs) for non-persistent compounds measured in single spot urine samples during pregnancy. ICCs tend to be higher for metabolites that come from personal care products or materials found in the home (e.g. MEP, MBzP) than for metabolites for which the likely source of exposure is dietary (e.g. BPA, DEHP) (199-202).”)

9. In section 5.2, it would be worthwhile to mention efforts to create international fetal growth standards (such as Intergrowth-21st) and to comment on the challenges of interpreting studies that might rely on different growth curves for these purposes.

RESPONSE: We have added several sentences mentioning the choice in using population-specific reference curves vs. customized growth curves vs. “universal standards” such as INTERGROWTH-21st at lines 1521-1528 (“Another consideration in the analysis of ultrasound data is the approach for calculating standardized measurements (i.e., z-scores or centiles) for each measurement. Most studies apply population-specific references (e.g., the Generation R cohort, the LIFECODES birth cohort, and the INMA cohort) (101, 220-222). However, alternative approaches, such as using customized growth curves (e.g., Buck-Louis et al. described above) or universal growth curves (e.g., INTERGROWTH-21st) are also options. While it is not clear what impact this choice has on associations between environmental exposures and fetal growth, this is a question worth investigating (223, 224).”)

Discretionary revisions, formatting, and typos

10. I would prefer to see the authors spell out expectations as opposed to speculating about typical expectations of the readers at lines 112-114 "These associations may interact with fetal
growth in the opposite direction that is typically expected but need to be carefully considered in the study of these contaminants and fetal growth."

RESPONSE: We have changed this sentence to read “Thus, researchers should be attentive to the potential for overgrowth of the fetus in response to chemical exposures as well.” (New lines 187-188).

11. At lines 210-218 the authors mention effects on fetal hormones, but don't explicitly mention the possible influence of EDCs on fetal/newborn thyroid hormones. This literature is decidedly mixed, but several of the EDCs of interest have been suggested to influence neonatal thyroid hormones. Dr. Tom Zoeller has published a few nice review articles on the topic.

RESPONSE: We have added a sentence to this section mentioning that EDCs also have the potential to influence fetal/newborn thyroid hormone levels (lines 150-152, “Phthalates, environmental phenols, and many pesticides fall under the classification of Endocrine Disrupting Compounds (EDCs) because of their ability to interfere with hormones (43).”)

12. The description of how studies were selected for the review in the methods is clear and very helpful. Though not absolutely necessary, a flowchart might be useful to help readers follow the process for selection of papers to be discussed.

RESPONSE: We thank Reviewer 1 for the feedback. As this review already has several tables and a figure, we are choosing not to add an additional figure to the paper.

13. The in text references are sometimes inside of and sometimes outside of the punctuation. I believe that inside of the punctuation is correct, but please consult journal guidelines to confirm.

RESPONSE: We thank Reviewer 1 for pointing out this inconsistency, and we have edited the formatting according to Vancouver reference style required by the journal.

14. Line 633 "number of benefits" - typo?

RESPONSE: We have edited this line accordingly.

15. Line 762 - parental◊ paternal - typo?

RESPONSE: We have edited this line accordingly.
Reviewer #2

This manuscript presents a detailed review of the literature examining associations of non-persistent chemicals and fetal growth as measured at birth and in utero. This is a welcome review of the literature given the increasing velocity of research on effects of these chemicals on perinatal outcomes. The authors synthesized a large literature base, and the review will be a valuable resource for those interested in the topic. The authors find that there is little consistency of findings and suggest several potential sources of study-to-study variation that may contribute to this heterogeneity. While this review is comprehensive, it could be improved by addressing several additional points.

Major comments:

1) This review paper compelling demonstrates that variability in the exposure and outcome assessment and statistical approaches likely leads to inconsistency of published results. However, the value and impact of this review could be greater if the authors made concrete recommendations to aid investigators planning new studies or evaluating the existing literature. Are there best practices for exposure assessment, outcome assessment, or statistical approaches that can be recommended? Are there specific areas in need of methods development? Specific examples of instances when recommendations would be useful are noted in the comments below.

RESPONSE: Thank you for these comments. While we agree that it would be wonderful to establish these best practices, these recommendations are difficult to make and need to be tailored to specific research questions (even within the overall question of the whether a compound is associated with fetal growth), chemicals of interest, and hypothesized mechanisms at work. However, we have made several recommendations in the new version of the review.

In regard to exposure assessment, we have added additional discussion of considerations in regard to type of urine sample collection (e.g., spot vs. 24-hour, lines 1042-1055), pooling of specimens (lines 1068-1194), season of exposure (1445-1450), and the need to assess windows of vulnerability to exposure (lines 1185-1189, 1207-1211) as well as the following specific lines:

Lines 1057-1058 “Because of this variability, measurement of exposure biomarkers in multiple specimens collected across pregnancy is recommended.”
“Measuring more than one sample of urine collected at different times of day, particularly relative to timing of a participant’s most recent meal or urination, can improve exposure characterization of chemicals with dietary sources (199, 200).”

“Thus, in an ideal setting, repeated urine samples would be collected across gestation and analyzed individually. Since this is not always feasible, alternative approaches—such as exploring windows of vulnerability in a subset and then subsequently pooling—are encouraged. Investigators should carefully consider the time period of exposure that one or more biomarkers reflect, as well as hypotheses regarding mechanisms of action when designing exposure assessment methods for large studies.”

“In summary, careful consideration in study design must be given to determining the mode of urine sample collection, number of specimens, and whether or not to pool. Striking a balance between cost, participant burden, and scientific integrity can be challenging in this field.”

In regard to outcome assessment, we have made it more clear that we support the use of ultrasound data to assess fetal growth meaningful, and specified that collecting multiple ultrasounds in the second half of gestation may be particularly important since this window is when most of the growth occurs (lines 1510-1519, “The lack of similarity in timing of ultrasound measurements makes comparing study results challenging. It may be particularly important in research studies of fetal growth to capture at least two measurements from the second half of pregnancy, when the most growth occurs. Our previous work has demonstrated that ultrasound measures taken later in pregnancy may be the most relevant for capturing associations with phthalate and phenol exposure (102, 149).”)

In regard to statistical approaches, we have described one particular unanswered question, which is how different methods for standardizing fetal growth (e.g., customized growth curves, universal standards, population references) may influence results (lines 1521-1528 “Another consideration in the analysis of ultrasound data is the approach for calculating standardized measurements (i.e., z-scores or centiles) for each measurement. Most studies apply population-specific references (e.g., the Generation R cohort, the LIFECODES birth cohort, and the INMA cohort) (101, 220-222). However, alternative approaches, such as using customized growth curves (e.g., Buck-Louis et al. described above) or universal growth curves (e.g., INTERGROWTH-21st) are also options. While it is not clear what impact this choice has on associations between environmental exposures and fetal growth, this is a question worth investigating (223, 224).”
Other additions described below in response to specific comments below.

2) Please include the reference numbers next to the study names in the Tables. Currently, it is extremely cumbersome to link information in the text to the data presented in the tables.

RESPONSE: Thank you for this suggestion, we have added study numbers to the tables.

Abstract

3) Here and throughout, consider replacing the term "phenols" with "environmental phenols, parabens, and organophosphate ester flame retardants" or "environmental phenols and other non-persistent chemicals": the scope of the review is on phenols arising from environmental sources and it does not include other phenolic compounds. Parabens and organophosphate ester flame retardants are not phenolic compounds.

RESPONSE: We have edited the lines referring to “phenols” in the abstract and throughout the manuscript and tables using the worded suggested by Reviewer 2.

4) Line 9: Is instability the right word here? Perhaps high temporal variability? Instability made me think of sample degradation or other laboratory factors.

RESPONSE: We have edited this word according to Reviewer 2’s suggestion.

Introduction

5) Lines 94-105: Would sex hormones also be relevant in this discussion? It seems they would be particularly important for phthalates, which are anti-androgens.

RESPONSE: We agree and have added estrogen/androgen hormones to the list in this section, but have not added any specific discussion about how these factors relate to fetal growth since the relationships are more tenuous than what is known about thyroid hormones and neuroendocrine systems (new line 172).

Methods

6) Line 231: What outcomes are meant by "whatever you measured"?

RESPONSE: This was a typo and has been deleted.
7) Lines 238-253: It appears that child's sex and, for the OP pesticides, PON are additional criteria that could be added to this section. Perhaps also assessment of non-monotonicity if this was done systematically?

RESPONSE: We have added wording in new lines 409-412 to describe these criteria (“We present results stratified by sex or restricted to a single sex, as the effects of prenatal exposure to some non-persistent environmental chemicals may differ by fetal sex. For studies of organophosphate pesticides, we additionally present results stratified by PON1 genotype and status.”). Non-monotonicity was not examined in a systematic manner, so we will not describe it in this section nor in the section on environmental phenols.

Results

8) Overall, I found this section difficult to evaluate since the references are by number in the text but by author in the Table.

RESPONSE: Thank you for the feedback. The journal guidelines indicate that the numbered Vancouver reference style is required, but we have added numbers to the tables to improve evaluation of the results.

9) The discussion of phthalate metabolites could be improved by considering individual metabolites or molecular weight groupings since biological activity varies among members of this large class. For example, in lines 284-290, which phthalates were associated with increased birth size and which with decreased? Was there a pattern by metabolite or molecular weight group?

RESPONSE: We agree and have added descriptions of individual phthalates and molecular weight groupings at lines 484-485 (“There were no notable patterns by phthalate metabolite or molecular weight, timing of exposure assessment, or outcome measured”), lines 492-510 (“Although two studies reported some inverse associations between some high molecular weight phthalate metabolites and birth weight or length in boys (98, 100), five others found positive associations between both low and high molecular weight urinary phthalate metabolites measured at different time points in pregnancy and birth size in boys (97, 101, 104, 109, 112). Among girls, concentrations of high molecular weight phthalates generally had null (100, 112, 113) or positive associations with birth size (101, 104, 109).”), lines 521-527 (“A study of 3100 births in China found that low molecular weight phthalate metabolites monomethyl phthalate (MMP) and monoethyl phthalate (MEP) were associated with reduced birth weight in the overall cohort and with birth length in girls (111). On the other hand, monobenzyl phthalate (MBzP) was..."
positively associated with birth weight in two studies (95, 101), and metabolites of dibutyl phthalate (DBP) were positively associated with birth weight in models restricted to boys (101, 109), restricted to girls (101), and overall (95).”), lines 580-583 (“Two small hospital-based case-control studies from the same research group reported that levels of DEHP metabolites measured at a single time point in the third trimester were associated with increased odds of IUGR or “fetal growth restriction” (diagnosis of either IUGR or low birth weight) (100, 126).”), lines 586-606 (“Urinary concentrations of high molecular weight phthalate metabolites – including MBzP, monocarboxy-isononyl phthalate (MCNP), and metabolites of DEHP – were statistically significantly inversely associated with both biparietal diameter and estimated fetal weight throughout pregnancy. MCNP, however, was significantly positively associated with ultrasound measures of femur length during gestation”), lines 609-612 (“While they found inverse associations between mono-n-butyl phthalate (MnBP, a metabolite of DBP) and fetal size and growth rates early in pregnancy (at and between 12 and 20 weeks gestation), they report positive associations between MBzP and MnBP and the rate of fetal growth between 20 and 34 weeks of gestation”), and lines 616-620 (“Although phthalate metabolite concentrations were not significantly associated with birth weight, cumulative exposure to high molecular weight phthalate metabolites (notably MBzP and metabolites of DEHP) over pregnancy was significantly negatively associated with head circumference, abdominal circumference, femur length, and estimated fetal weight. MEP was associated with reduced head circumference in female fetuses only.”)

10) Lines 301-307: It is unclear why tests of non-monotonicity are discussed only for these two studies. Did no other studies assess non-linear dose response relationships, or did you choose to only report on studies for which there was evidence of non-monotonicity?

RESPONSE: We thank Reviewer 2 for pointing out this lack of clarity. We had chosen to note two studies that had highlighted evidence of non-monotonic dose-response curves in their results. Rather than imply that non-monotonicity was examined in a systematic manner, we are consolidating comments about non-linear or non-monotonic dose-response curves to the discussion section 5.3 Statistical approaches and biases (specifically, lines 1552-1561, “These models assumed a monotonic, if not linear, relationship between exposures and outcomes. However, a number of studies that examined categories (tertiles or quartiles) of prenatal phthalate levels found few monotonic trends but identified non-monotonic statistically significant associations (96, 147, 152). If physiological responses to these chemicals exist on a non-linear dose-response curve, it is possible that continuous linear regression models may be unable to detect real effects. We therefore recommend investigators examine non-linear and non-monotonic dose response curves. While categorical exposure variables are both simple to create and easy to interpret, they can be subject to limitations (228, 229). Flexible approaches to assessing dose-response relationships, such as nonparametric regression, fractional polynomial
regression, or the use of splines, may further improve assessment of the shape of dose response curves (228).”) See also: response to Reviewer 2 comment #22.

11) Lines 328-355: These paragraphs state the information found in the table. This text could be condensed or replaced by a synthesis of whether the five papers in Table 1B are consistent (as was done for Table 1A). In particular, the conclusions noted in the Summary section (4.1.3) are highly relevant but not fully discussed/supported in the preceding sections.

RESPONSE: We thank Reviewer 2 for these suggestions and have added a brief summary of Table 1B studies at lines 576-578 (“These studies varied by size (from 119 to 520 infants), timing and number of urine samples collected, phthalate metabolites measured, outcomes assessed, covariates included in multivariable models, statistical methodology, and associations reported”), condensed the description of studies at lines 580-583 (“Two small hospital-based case-control studies from the same research group reported that levels of DEHP metabolites measured at a single time point in the third trimester were associated with increased odds of IUGR or “fetal growth restriction” (diagnosis of either IUGR or low birth weight) (100, 126).”), and added detail regarding specific phthalates and studies with multiple measures of phthalates (referenced in Summary section 4.1.3) at lines 586-620 (see also: response to Reviewer 2 comment #9).

12) Line 372: This statement warrants a citation.

RESPONSE: We have edited this line accordingly.

Discussion

13) Lines 593-601: How would conflicting results lead one to believe that human pregnancy is impervious to the exposures? This argument would hold if all studies were well-powered and null, but many of the studies in the table reported associations.

RESPONSE: The point we were making here is that because all studies were not well-powered and null, as Reviewer 2 points out, we should not conclude that there is no effect of these chemicals on fetal growth. We thank Reviewer 2 for pointing out why this section could be interpreted otherwise and have edited new lines 1008-1111 to clarify this point (“However, the epidemiologic evidence for such associations is inconsistent. Because relatively few studies exist, and because of the particular complexity in evaluating the relationships between non-persistent chemicals and fetal growth, we believe that many current studies are insufficiently powered or inadequately designed to detect effects.”).
14) Lines 593-601: Beyond study design, it seems that there are a number of additional considerations when evaluating "conflicting conclusions" - such as differences in exposure level and population susceptibility that could result in differences in true effects between studies - these are not discussed. In particular, exposure level could be an important factor explaining differences between studies. While reporting these levels for each chemical and study is probably beyond the scope of this paper, perhaps some discussion of exposure levels could be incorporated in the text? For example, I wondered whether dichlorophenol levels were lower in the one study that did not find an association compared to the four that did (Line 424-426).

RESPONSE: We agree with Reviewer 2 that exposure levels and population susceptibility could contribute to different true effects and that reporting exposure levels for every chemical in every study is beyond the scope of this review. The focus of the limitations of this review is in variability in methodology, as these are areas that researchers can actively change (whereas population exposure levels and susceptibility can be described, but not changed, by researchers). We have added lines 1352-1450 to highlight this point and to recommend that researchers compare the exposure levels in their study populations to those in other studies and in population-based samples such as NHANES ("Finally, variation in levels of exposure to non-persistent environmental chemicals, as well as differences in the susceptibility of populations under study, can contribute to differences in the true effect between studies of the same exposure and outcome. Reporting the concentrations of every chemical measured in every study described is beyond the scope of this review. Rather, we recommend that researchers compare chemical exposure levels in their study population to those in both other study populations and in population-based samples (such as the National Health and Nutrition Examination Survey in the U.S.) to facilitate evaluation of these possible differences. Moreover, there is evidence that exposure to some non-persistent chemicals – such as BP3, TCS, and organophosphate pesticides – follow seasonal patterns (209-212). Birth weight also follows seasonal patterns, though these patterns can vary by population and years under study (213-215). Researchers should critically evaluate (using, for example, directed acyclic graphs (216)), whether season of measurement should be considered in modeling effects of non-persistent chemicals on fetal growth outcomes."). We have also added a note about dichlorophenol levels in the results section at line 733 ("despite similar distributions of dichlorophenol concentrations").

15) Line 603: Reading through the results, there is an implicit suggestion that the authors prefer studies that use repeated exposure measures (Line 314, Line 408, Line 413, Line 492), though the authors' stance is less clear when this issue is discussed in Section 5.1. While I don't disagree that repeated measures are useful, there is an open question of whether averaging exposure over all of pregnancy is always better for analyses of health outcomes: if the critical window of susceptibility were in the first trimester, averaging levels across the first, second, and third
 trimester may actually introduce exposure misclassification compared to the single first trimester measure. Is it known whether there is a susceptible period for the fetal growth outcomes discussed in this paper? I did not see a discussion of susceptible periods in the background section. Related, it would also be important to note if there is no evidence of susceptible periods - if cumulative exposure is indeed most important, this supports averaging exposures over pregnancy.

RESPONSE: Thank you for this response, we agree that this section warrants a discussion of susceptible periods. We have adding sentences to address this topic lines 1185-1190 (“An important limitation to pooling samples across weeks of pregnancy, however, is that key windows of vulnerability to exposure may be missed. Consider a chemical for which exposure during the first trimester is the most relevant for fetal growth and for which there is high variability (low ICC) across pregnancy. If the urine sample from this time point is pooled with those collected later in pregnancy, any potential associations would be diluted.”), and lines 1209-1217 (“Investigators should carefully consider the time period of exposure that one or more biomarkers reflect, as well as hypotheses regarding mechanisms of action when designing exposure assessment methods for large studies. It should be noted that the windows of exposure measured were highly variable across the literature reviewed here. If the growth of the fetus is more vulnerable to environmental stressors during one point in gestation than another, this variation likely contributes to the lack of consistency seen in results. While we did not formally evaluate whether associations were more consistent when biomarkers were measured earlier versus later in pregnancy, we observed no clear patterns in associations by timing of exposure assessment.”). However, the period(s) of greatest susceptibility to environmental perturbations may depend on the chemical and/or the mechanism of effect. We therefore avoid specific recommendations, but rather added points for researchers to consider as they approach this topic (lines 1206-1209 “Thus, in an ideal setting, repeated urine samples would be collected within trimesters and across gestation and analyzed individually. Since this is not always feasible financially, alternative approaches—such as exploring windows of vulnerability in a subset and then subsequently pooling—are encouraged.” and lines 1219-1221 “In summary, careful consideration in study design must be given to determining the mode of urine sample collection, number of specimens, and whether or not to pool. Striking a balance between cost, participant burden, and scientific integrity can be challenging in this field.”).

16) Line 603: Do you have a recommendation about the number and timing of repeated measures? Perhaps repeated measures within each trimester would allow for improved exposure assessment without compromising ability to assess susceptible periods? It may also be worthwhile to discuss potentials solutions to this issue, such as leveraging pooled samples or statistical methods for exposure measurement error correction (see Perrier et al. Epidemiology 2016; PMID: 27035688).
RESPONSE: We have included additional recommendations to improve exposure assessment at lines 1062-1065 (“Measuring more than one sample of urine collected at different times of day, particularly relative to timing of a participant’s most recent meal or urination, can improve exposure characterization of chemicals with dietary sources (199, 200).”), lines 1052-1058 (“Because of this variability, measurement of exposure biomarkers in multiple specimens collected across pregnancy is recommended.”), lines 1206-1209 (“Thus, in an ideal setting, repeated urine samples would be collected within trimesters and across gestation and analyzed individually. Since this is not always feasible financially, alternative approaches—such as exploring windows of vulnerability in a subset and then subsequently pooling—are encouraged. Investigators should carefully consider the time period of exposure that one or more biomarkers reflect, as well as hypotheses regarding mechanisms of action when designing exposure assessment methods for large studies.”).

We have also added a discussion of within-subject pooling of biospecimens at lines 1067-1189 (“Researchers may be reluctant to measure numerous biomarkers during pregnancy due to high cost of laboratory assays. Within-subject pooling of biospecimens, where samples from a single individual at multiple time points are combined prior to measurement, can be used to address this concern while also reducing misclassification of exposure assessment (203). Increasing the number of biospecimens in an individual’s pooled assay can both decrease bias in the effect estimate and increase power (203). Additionally, within-subject pooling can improve exposure characterization over first morning voids (201). At least 6 and 35 specimens are required to limit bias to 10% attenuation for chemical with ICC of 0.6 and 0.2, respectively, though (203). This number of biospecimens may be unfeasible to collect for logistical or financial reasons. However, if the same number of biospecimens are pooled for each participant, and reliable estimates of ICCs are available, a posteriori disattenuation correction can virtually eliminate bias in effect estimates (203). Moreover, if at least two biospecimens are measured separately, measurement error models such as simulation extrapolation or regression calibration can be used to reduce bias to less than 10% (203). An important limitation to pooling samples across weeks of pregnancy, however, is that key windows of vulnerability to exposure may be missed. Consider a chemical for which exposure during the first trimester is the most relevant for fetal growth and for which there is high variability (low ICC) across pregnancy. If the urine sample from this time point is pooled with those collected later in pregnancy, any potential associations would be diluted.”)

17) Line 603: I wondered if the authors' have a recommendation regarding type of urine sample? Were there differences in associations based on whether spot urines, first morning voids, or 24-hour urines were analyzed? There are major differences in how well each of these types of samples perform for different chemicals based on exposure sources.
RESPONSE: We did not systematically evaluate whether there were differences in associations detected by type or time of day of urine sample collection. However, we have added a brief discussion of types of urine samples at lines 1042-1055 (“Thus, relying on a single spot urine measurement of a non-persistent chemical can induce bias in its estimated effect, with as much as 40% attenuation in the effect estimate even with an ICC as high as 0.60 (203). Measuring concentrations of a chemical in a 24 hour urine sample is more representative of the day’s exposure compared to a spot urine sample (204, 205). First morning void samples are more complicated because time of day is a significant predictor of levels of phthalates and BPA in urine, with higher levels of BPA and high molecular weight phthalates observed in samples collected in the evening, and highest levels of MEP in the morning (200, 201, 206).”).

18) Lines 619-621: Considering this suggestion that timing of exposure may be important, did you evaluate whether associations were more consistent among studies that assessed biomarkers in early versus late pregnancy?

RESPONSE: Although we did not evaluate whether associations were more consistent among studies that assessed biomarkers in early versus late pregnancy, though we did not notice any clear patterns of associations by timing of exposure assessment. We have added lines 1211-1217 (“It should be noted that the windows of exposure measured were highly variable across the literature reviewed here. If the growth of the fetus is more vulnerable to environmental stressors during one point in gestation than another, this variation likely contributes to the lack of consistency seen in results. While we did not formally evaluate whether associations were more consistent when biomarkers were measured earlier versus later in pregnancy, we observed no clear patterns in associations by timing of exposure assessment.”) to indicate so in the text.

19) Lines 622-628: This argument about reproducibility of low versus high molecular weight phthalate concentrations due to differences in exposure sources would be strengthened by reporting quantitative information from the literature, such as correlation coefficients or ICCs.

RESPONSE: We agree and have added a discussion of intraclass correlation coefficients, with appropriate references, to new lines 1037-1042 (“This is exemplified by an extensive literature on intraclass correlation coefficients (ICCs) for non-persistent compounds measured in single spot urine samples during pregnancy. ICCs tend to be higher for metabolites that come from personal care products or materials found in the home (e.g. MEP, MBzP) than for metabolites for which the likely source of exposure is dietary (e.g. BPA, DEHP) (199-202).”)
20) Lines 603-628: Considering these points, what are your recommendations about measurement of urinary biomarkers of non-persistent chemicals during pregnancy? There is a solid recommendation regarding temporality, but are there other "best practices" you would suggest?

RESPONSE: We have added more detail to our discussion of the benefits and limitations of the variety of options available in collecting and analyzing urinary biomarkers of non-persistent chemicals during pregnancy (e.g. spot v. 24-hr urine measurements and pooling specimens). We have also added a few additional specific recommendations at lines 1057-1058 (“Because of this variability, measurement of exposure biomarkers in multiple specimens collected across pregnancy is recommended.”), lines 1063-1065 (“Measuring more than one sample of urine collected at different times of day, particularly relative to timing of a participant’s most recent meal or urination, can improve exposure characterization of chemicals with dietary sources (199, 200)”), lines 1207-1211 (“Since this is not always feasible financially, alternative approaches—such as exploring windows of vulnerability in a subset and then subsequently pooling—are encouraged. Investigators should carefully consider the time period of exposure that one or more biomarkers reflect, as well as hypotheses regarding mechanisms of action when designing exposure assessment methods for large studies.”), lines 1441-1444 (“Rather, we recommend that researchers compare chemical exposure levels in their study population to those in both other study populations and in population-based samples (such as the National Health and Nutrition Examination Survey in the U.S.) to facilitate evaluation of these possible differences.”), and lines 1448-1450 (“Researchers should critically evaluate (using, for example, directed acyclic graphs (216)), whether season of measurement should be considered in modeling effects of non-persistent chemicals on fetal growth outcomes.”).

21) Lines 681-686: If the lack of similarity in timing of ultrasound measurements is a limitation of the current literature base, do you have a recommendation for when these measurements should be conducted?

RESPONSE: We have added a note specifying that ultrasounds from the second half of pregnancy may be preferred, as that is the time period when the most growth occurs (Lines 1511-1519, “It may be particularly important in research studies of fetal growth to capture at least two measurements from the second half of pregnancy, when the most growth occurs. Our previous work has demonstrated that ultrasound measures taken later in pregnancy may be the most relevant for capturing associations with phthalate and phenol exposure (102, 149).”).

22) Line 708-710: Consider recommending splines or other flexible approaches for assessing non-monotonic dose-response relationships given that using categorized exposure variables (i.e., quantiles) is itself subject to limitations (see Greenland Epidemiology 1995; PMID: 7548341).
RESPONSE: While categorical exposure variables are subject to limitations, these limitations may not be as severe as suggested by Greenland 1995, as described in a response by Weinberg 1995 (PMID: 7548338). We agree, though, that more specific recommendations should be included, which have been added at lines 1156-1561 (“We therefore recommend investigators examine non-linear and non-monotonic dose response curves. While categorical exposure variables are both simple to create and easy to interpret, they can be subject to limitations (228, 229). Flexible approaches to assessing dose-response relationships, such as nonparametric regression, fractional polynomial regression, or the use of splines, may further improve assessment of the shape of dose response curves (228).”)

23) Lines 740-741: If there are differences in interpretation between methods, do you have a recommendation as to what statistical approach(es) should be used to assess modification by sex?

RESPONSE: We have added wording at lines 1633-1639 to recommend the augmented product approach described by Buckley et al. 2017 (“We recommend an alternative augmented product term approach described by Buckley et al., which entails including both an exposure by sex product term and product terms for covariates by sex (250). This method produces the same effect estimates as stratification but allows for formal statistical evaluation of heterogeneity using a Wald test or likelihood ratio test of the exposure by sex product term (250). In this area of research, examination of sex differences should be standard, and methods for investigating those differences clearly relayed.”)

24) Line 697: I expected to see some discussion of diet as a confounder in this section. Maternal diet during pregnancy is a primary source of exposure to many non-persistent chemicals and is also strongly related to fetal growth, but often not well-characterized in environmental epidemiology studies. Do the authors have an opinion on whether confounding by diet may explain differences in results among studies, or whether information on maternal diet is a key confounder that should be addressed?

RESPONSE: Thank you for this suggestion, we have added a discussion of diet as a confounder to this section at lines 1605-1618 (“Maternal diet during pregnancy influences fetal growth and is also a primary source of exposure to some non-persistent chemicals (238). Increased caloric intake during pregnancy is associated with increased birth weight (239), although there is evidence that consuming a diet high in processed or red meat, or high fat dairy, during pregnancy is associated with increased odds of giving birth to an SGA infant (240). Eating canned food, fish, and fast food have also been shown to be positively correlated with BPA levels in pregnant women (241-243), and other bisphenols, such as BPS, have been detected in food as well (244). Poultry, high-fat dairy, and fast food consumption may all be sources of exposure to phthalates
such as DEHP (91, 245). Dietary factors are often not well-characterized in environmental epidemiology studies and likely confound the relationship between prenatal exposure to some non-persistent chemicals and fetal growth. The limited or nonexistent control for these factors in statistical models or study design could explain some of the variability in the results among studies of chemicals for which diet is a primary source of exposure. Careful evaluation of these entangled relationships is therefore warranted.”