PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	The Importance of Maternal Diet Quality During Pregnancy on Cognitive and Behavioural Outcomes in Children – A Systematic Review and Meta-Analysis
AUTHORS	Borge, Tiril; Aase, Heidi; Brantsaeter, Anne-Lise; Biele, Guido

VERSION 1 - REVIEW

REVIEWER	Russell de Souza and Michael Zulyniak McMaster University, Hamilton, ON, Canada
REVIEW RETURNED	12-Aug-2016

GENERAL COMMENTS	This is an exceptionally well conducted and written systematic review and set of meta-analyses. The objectives are clearly stated, methods described clearly, and the interpretation of the results is careful. We have three suggestions that relate to methodology: 1. The authors correctly draw attention to the fact that the list of covariates included in multivariable models in the included sudies are diverse. Thus, I would suggest presenting a meta-analysis of the minimally-adjusted model as well as the most-adjusted models. This would allow a reader to gauge how serious an impact the different methods of adjustment might have on the results of the meta-anaylsis. 2. The authors detail that their inclusion criteria was loosened to allow for the inclusion of more studies that addressed the research question. They list fish, n-3:n-6 ratio, and dietary fibre as good proxies for diet quality--and support this with references. Further, they authors note correctly that fibre-rich foods are part of tools such as MedDiet index, and HEI-2010. This raised to me the question of whether studies that used these pattern scores (e.g. MedDiet index or HEI-2010) were considered for inclusion in the meta-analysis. If so, were none found? If not, why were these not considered for the review? 3. Finally, the limitations and harmonization of dietary assessment was presented clearly; however, the harmonization of neurocognitive assessment questionnaires was presented in less detail. While the use of mean effect sizes to standardize measurement units between assessment methods was discussed, it would first be useful to know how similar/comparable each of the tests are for each of the outcomes. For example, according to Table 2, it appears that 'cognitive development' was measured using WISQ-III, MCDI, WASI, SDQ, and BSID-II. Are these tests all directly comparable? Please explain.

REVIEWER	Edward D Barker King's College London. UK
REVIEW RETURNED	22-Aug-2016

GENERAL COMMENTS	Interesting and very much needed meta-analysis. There are a lot of things to like about this paper including that it is well conceptualized and well written. The analyses appear appropriate as well, though I have no training in meta-analyses, so I will defer to the reviewers with appropriate knowledge. I have a few comments that can hopefully help improve an already impressive piece of work. 1. Re the discussion about breastfeeding, might the authors also mention that it can be a risks, esp for obese mothers (recent research published on this). 2. Moderators: Might sex of child be a moderator? I do believe there is talk about prenatal nutrition possibly affecting male children more than female. 3. Controlling for child diet. I do believe that Barker et al (2013) did control for postnatal child diet.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1
6. The authors correctly draw attention to the fact that the list of covariates included in multivariable models in the included studies are diverse. Thus, I would suggest presenting a meta-analysis of the minimally-adjusted model as well as the most-adjusted models. This would allow a reader to gauge how serious an impact the different methods of adjustment might have on the results of the meta-analysis.
Answer: We attempted to do this, however only three of the included articles provide information on the unadjusted effect measures. The remaining articles provide either only the fully adjusted effect sizes or partially adjusted as well as the fully adjusted effect sizes. We have now included the following clarification regarding this in the revised manuscript (3rd paragraph under heading "Summarizing effect sizes"): "We initially considered including the corresponding unadjusted effect sizes for each study, however only four studies provided this information [29, 31, 37, 38], and the studies that reported minimally adjusted results adjusted for different variables."

7. The authors detail that their inclusion criteria was loosened to allow for the inclusion of more studies that addressed the research question. They list fish, n-3:n-6 ratio, and dietary fibre as good proxies for diet quality--and support this with references. Further, they authors note correctly that fibre-rich foods are part of tools such as MedDiet index, and HEI-2010. This raised to me the question of whether studies that used these pattern scores (e.g. MedDiet index or HEI-2010) were considered for inclusion in the meta-analysis. If so, were none found? If not, why were these not considered for the review?
Answer: We were aware of these indices and assumed that the broad search strategy we used would have identified studies using them. To confirm this, we explicitly searched for studies that employed these indices to investigate the effect of maternal nutrition on child cognitive development. We did not identify any such studies. However, we appreciate that we probably should have explicitly included the name of the indices in the original search string to avoid any confusion, and have therefore included the following text in the revised manuscript (1st paragraph under Results): "Additionally, we explicitly searched for relevant studies that employed a priori dietary indices (like the HEI-2010 and Mediterranean diet index). No new relevant articles were retrieved."

8. Finally, the limitations and harmonization of dietary assessment was presented clearly; however, the harmonization of neurocognitive assessment questionnaires was presented in less detail. While the use of mean effect sizes to standardize measurement units between assessment methods was discussed, it would first be useful to know how similar/comparable each of the tests are for each of the outcomes. For example, according to Table 2, it appears that 'cognitive development' was measured using WISQ-III, MCDI, WASI, SDQ, and BSID-II. Are these tests all directly comparable? Please explain.
Answer: We attempted to group the outcome measurements based on what overarching domain each instrument where measuring. Ideally, one would prefer to separate the outcomes based on being measured by identical instruments, but considering the limited amount of studies relevant for inclusion in the meta-analysis, this was not feasible. The grouping was done in conjunction with recent research indicating that language, cognition, and executive functions are more strongly correlated with each other than with affective functioning. The literature on child development often separates functional domains involving affective or emotional control from more neutral functional areas (intellectual functioning, working memory, language. We have revised Table 2 so that it becomes clearer which measures are grouped either as cognitive development or as affective functioning.
To further highlight this we have included the following text in the revised manuscript (1st paragraph under the heading Summarizing effect sizes): "Selection of the outcome measures into each respective domain was based on 1) a thorough review of the properties of each instrument with regards to what area of development the instrument is aimed at measuring, based on the manual for each instrument, and 2) research indicating that language, cognition, and executive functions are more strongly correlated with each other than with affective functioning.[46, 47]"

Reviewer 2
9. Re the discussion about breastfeeding, might the authors also mention that it can be a risk, esp for obese mothers (recent research published on this).
Answer: Thank you for this input. In the revised version we have now mentioned that obesity is a factor that further complicates the role of breastfeeding as a confounding factor, as obesity is linked both to reduced breastfeeding as well as impaired cognition and behavior in offspring. The following text has been added following the discussion about breastfeeding (last paragraph before Limitations): "Maternal obesity is another factor related to both breastfeeding and the outcomes. Studies consistently show reduced breastfeeding rates in obese mothers[75, 76] and obesity is also linked to impaired cognition and increased behavioral problems in children.[77-79] This highlights the importance of also accounting for maternal BMI in the analysis.".
10. Moderators: Might sex of child be a moderator? I do believe there is talk about prenatal nutrition possibly affecting male children more than female.
Answer: Thank you for highlighting this interesting point. Gender might definitely be a moderator, and the majority of the studies adjusted for gender in their analysis, as seen from table 3, even though none of the studies particularly mentioned whether gender did indeed influence the results or not. Since all studies included both genders in their sample, as well as adjusting for gender in their analysis, it was not feasible to consider gender as a moderator for the purpose of this meta-analysis.

11. Controlling for child diet. I do believe that Barker et al (2013) did control for postnatal child diet.
Answer: We are not able to find any reference to postnatal child diet as a control variable in the Barker et al (2013) study. However, the authors controlled for postnatal maternal diet and also for contextual risk factors including housing condition, poverty etc. (points a-g on page 418), which are closely related to child diet. Other publications from the ALSPAC cohort have shown that throughout childhood, children and parents have similar dietary patterns and that their diet quality varies with sociodemographic variables (Emmet, Jones and Northstone, 2015). Hence, controlling for maternal postnatal diet and contextual risk factors is likely to also control for child diet, and the results clearly showed that maternal prenatal diet was the key exposure in this study. To further highlight the important association between maternal and child diet, we have revised the 1st paragraph under the

heading Maternal diet – direct effect or marker for child diet? as follows: “An important issue that cannot be resolved by this meta-analysis is whether the observed association is based on true effects stemming from maternal diet, or whether maternal diet is just a marker for the child’s diet. It is recognized that food preferences and eating habits established in early childhood generally forms the foundation for diet quality in adult life.[62, 63] Not surprisingly, maternal diet during pregnancy,[63, 64], as well as maternal post-natal diet,[65] and child diet during infancy and early toddlerhood has been found to be highly correlated which indicates that it is crucial to control for child diet when investigating effects of maternal diet during pregnancy (and vice versa). Also, in addition to being correlated, maternal and child diet quality varies with sociodemographic variables [65]. Hence, controlling for maternal postnatal diet and sociodemographic risk factors is likely to, at least in part, control for child diet, however child diet should ideally be assessed as a distinct factor.”

ADDITIONAL REVISIONS:
ADDITIONAL REVISIONS:
I. Since submitting the original manuscript it has come to our attention that there is some debate in the literature regarding optimal methods of identifying and adjusting for publication bias in meta-analysis. We have therefore included some additional analysis in the revised manuscript, and a description of this has been included in both the methods and results section where publication bias is described.
II. Two short paragraphs have been moved as this was deemed appropriate; one from the discussion to implications and one from limitations to results. These changes are marked in the marked version of the revised manuscript.
III. After all revisions the manuscript was thoroughly proof read and some minor grammatical changes were made. These changes are not marked in the marked copy of the revised manuscript.

VERSION 2 – REVIEW

REVIEWER	Russell de Souza McMaster University, Hamilton, ON, Canada
REVIEW RETURNED	08-Oct-2016

GENERAL COMMENTS	This manuscript has been extensively revised to address many of the comments raised by me and other reviewers, and feels in many ways, like a de novo manuscript. Much of it is improved. However, I feel that some aspects of this revised manuscript require further clarification before it is published. There were some flaws that, upon reviewing my earlier review, I did not fully appreciate, and I hope that presenting them now is acceptable, in the spirit of improving the presentation and interpretability of the work that has been done. 1. The authors switch back and forth between singular and plural within the same sentence on several occasions. Please read carefully (e.g. page 3, lines 52-54: "since proxies for maternal diet quality was included") 2. Might the authors provide a clearer justification for this 1990 cutoff? The authors rightly claim that dietary pattern analysis is relatively new, but I note that they also considered single nutrients (as proxies of diet quality). Therefore, studies prior to 1990 that examined these same single nutrients are not excluded on strong grounds.

	3. Data collection and extraction: This needs more detail. The goal of this section should be to clearly state how the literature filtering process was completed. I have several questions after reading this. How many authors screened abstracts? If >1, were they screened independently and then cross-checked? How many authors screened full-text? If >1, were they screened independently and then cross checked? Was there a process in place for resolving disputes, if these were done in duplicate. What is meant by "the process of summarizing effect measures from each study were performed multiple times"? Was data abstraction performed independently, by two reviewers, in duplicate? 4. Please update the search through Oct. 1, 2016 at least. 5. Table 4: can you please provide the following additional information... 1) for each row of this table, provide the number of studies, the number of participants. 2) for the top rows "Affective" and "Cognitive", please clarify whether the top row bearing these headings are the sum of the 2 or 3 indented, specific domains beneath. 6. Moderator analysis (p. 11, line 19). What is meant by "a moderator analysis is generally not applicable"? Perhaps you mean is not advisable when there are <10 studies, to avoid risk of spurious findings? If so, please be explicit, with a reference, of why proceeding with moderator analysis in this scenario (<!0 studies) is ill-advised. 7. Moderator analysis (p..11, line 24-33). More methodological detail is required, either here or in "Methods". How was the moderator analysis conducted? What is a "single-predictor" model? Do you mean the variable was entered as an IV in a meta-regression model? I cannot understand from your description how this was carried out. Further, what is meant by "diet category"? How did you test for moderation of the observed "effect" (I prefer "association" for observational studies). 8. Discussion: "Maternal diet-- direct effect or marker for child diet?" -- I wonder if it is truly appropriate to control for child diet? If the hypothesis is that maternal diet --> child diet --> cognitive effects, then child diet would not be classical confounder. IT would be on the causal pathway from exposure (maternal diet) to outcome (cognitive effects in child). It is complicated, and this aspect was not apparent to me on readign this section. Perhaps I misread, but I encourage the authors to re-read this paragraph and ensure that this has been fully considered. 10. Overall, I prefer the term "association" rather than "effect" in this observational meta-analysis. 11. I am very surprised no effort was made to address study risk of bias? Publiication bias is one form of potential bias, but the individual studies need to be assess for factors that might have led to biased results. I do not buy the authors' justification that because there is no "causal interpretation" attempted, quality issues of the included studies are irrelevant. All observational studies are limited for causal inference, but a systematic reviewer must give his/her readers an assessment of the quality of the included studies. This either needs to be completed, or a very strong rationale for not doing so needs to be presented in the paper itself.

	12. Please confirm that in Figure 2, the ultimate diamond (0.13 [0.09, 0.17]) does not include results from the same study cohort, with the same measure in the same participants only at different times. For example, if Bernard et al 2013.1, Bernard et al 2013 each contained the same participants, then you are double counting, and the overall effect estimate effectively double weights this one study (Bernard 2013). I do appreciate the efforts the authors have taken to attempt to explain this, but it's still not as clear as it should be to be confident in the forest plot. 12. Was the search strategy developed with an information specialist/librarian? If so, please state this. If not, I suppose you can avoid mentioning it.

REVIEWER	Edward D. Barker, PhD Institute of Psychiatry, Psychology and Neuroscience, King;s College London
REVIEW RETURNED	03-Oct-2016

GENERAL COMMENTS	The authors have done a very nice job responding to reviewer comments. The manuscript is improved. I have no further comments.

VERSION 2 – AUTHOR RESPONSE

Reviewer Name: Russell de Souza

Institution and Country: McMaster University, Hamilton, ON, Canada Competing Interests: None declared.

1. The authors switch back and forth between singular and plural within the same sentence on several occasions. Please read carefully (e.g. page 3, lines 52-54: "since proxies for maternal diet quality was included")

ANSWER: After all revisions were made, we thoroughly read the manuscript with particular focus on grammar and tense to correct for the abovementioned linguistic errors.

2. Might the authors provide a clearer justification for this 1990 cut-off? The authors rightly claim that dietary pattern analysis is relatively new, but I note that they also considered single nutrients (as proxies of diet quality). Therefore, studies prior to 1990 that examined these same single nutrients are not excluded on strong grounds.

ANSWER: When considering this comment we agreed that we did not have an adequately strong reason for restricting the literature search to studies published from 1990 and onwards. We therefore performed an updated literature search without this time-restriction. No relevant articles prior to 1990 were retrieved.

3. Data collection and extraction: This needs more detail. The goal of this section should be to clearly state how the literature filtering process was completed. I have several questions after reading this. How many authors screened abstracts? If >1, were they screened independently and then crosschecked? How many authors screened full-text? If >1, were they screened independently and then crosschecked? Was there a process in place for resolving disputes, if these were done in duplicate? What is meant by "the process of summarizing effect measures from each

study were performed multiple times"? Was data abstraction performed independently, by two reviewers, in duplicate?

ANSWER: Thank you for pointing this out. We have revised the text below this heading to the following: “After identification of original articles from the initial search, excluding duplicates, TCB and ALB independently screened title and/or abstract of each study. TCB’s and ALB’s final list of eligible and possibly eligible studies were then crosschecked and read in full text by both. If both reviewers were unsure whether an article was eligible for inclusion, GB was consulted to assure coherence regarding the final selection of articles for inclusion. After identification of the eligible articles the relevant information from each study were extracted by TCB (e.g. year of publication, total number of participants, dietary exposure and outcome measures assessed, confounders controlled for, and reported effect sizes) in collaboration with GB and ALB. TCB and GB then assessed individual study quality with the Newcastle-Ottawa Scale (NOS) for assessing the quality of cohort studies in meta-analysis.”

4. Please update the search through Oct. 1, 2016 at least.

ANSWER: We performed an updated literature search Nov. 16, 2016, removing the publication year restriction criteria (1990-current). Four new eligible articles were retrieved, all published in 2016. Information on the updated literature search has been added under the “Study sample and selection” heading in the results section.

5. Table 4: can you please provide the following additional information... 1) for each row of this table; provide the number of studies, the number of participants. 2) For the top rows "Affective" and "Cognitive", please clarify whether the top row bearing these headings are the sum of the 2 or 3 indented, specific domains beneath.

ANSWER: The requested information has been added to the table.

6. Moderator analysis (p. 11, line 19). What is meant by "a moderator analysis is generally not applicable"? Perhaps you mean is not advisable when there are <10 studies, to avoid risk of spurious findings? If so, please be explicit, with a reference, of why proceeding with moderator analysis in this scenario (<!0 studies) is ill advised.

ANSWER: Thank you for this feedback. We apologize if the wording was unclear, as we indeed referred to the fact that moderator analysis potentially leads to spurious findings when the number of studies is small. We have revised the text accordingly: “Considering that performing a moderator analysis is generally not advisable with less than ten studies, we performed a moderator analysis for the whole sample of studies, rather than separately for the affective domain.”

7. Moderator analysis (p.11, line 24-33). More methodological detail is required, either here or in "Methods". How was the moderator analysis conducted? What is a "single-predictor" model? Do you mean the variable was entered as an IV in a meta-regression model? I cannot understand from your description how this was carried out. Further, what is meant by "diet category"? How did you test for moderation of the observed "effect" (I prefer "association" for observational studies).

ANSWER: The moderator analysis was performed by estimating meta-regression models, where potential moderator variables were added as additional independent variables to single-predictor models – i.e. each moderator was assessed in separate models.

Diet category refers to whether the respective effect size is based on assessment of dietary pattern or a proxy for a healthy diet (fish intake, fruit intake etc.) – this is explained in the paragraph under the heading “Possible moderators”. To include more detail about the moderator analysis, the paragraph concerning this analysis in the Methods section has been revised to the following: “If the REM analyses indicate presence of heterogeneity, moderator analyses can be used to investigate potential

causes of this heterogeneity. We used meta-regressions where we added potential moderator variables individually to separate regression models in order to assess their effect on the association between exposure and outcome. The following factors were available for consideration as possible moderators: Publication year, diet category (type of diet classification - whether the exposure is defined as maternal dietary pattern or a proxy for maternal dietary pattern) and instrument category (measurement of outcome - questionnaire or neuropsychological test), as they are all factors which might moderate the effect of the exposure. The categorical factors were dichotomous"

We also took care to be more precise in the use of the terms "effect", "effect size", "association", such that the term "effect" is now only used where we discuss causal effects. Otherwise we use the terms "effect size" or "association" when presenting and discussing the results.

8. Discussion: "Maternal diet-- direct effect or marker for child diet?" -- I wonder if it is truly appropriate to control for child diet? If the hypothesis were that maternal diet --> child diet --> cognitive effects, then child diet would not be classical confounder. IT would be on the causal pathway from exposure (maternal diet) to outcome (cognitive effects in child). It is complicated, and this aspect was not apparent to me on reading this section. Perhaps I misread, but I encourage the authors to re-read this paragraph and ensure that this has been fully considered.

ANSWER: We regret our imprecise use of the term confounder. In using this term, we were referring to an alternative exposure that is associated with both the exposure of interest and the outcome. We recognize that the term "confounder" typically refers to a variable that is an ancestor to both the exposure and the outcome. However, the key interest of our manuscript is the direct effect of maternal diet during pregnancy, rather than the total effect of maternal and child diet combined. To estimate this direct effect accurately it is necessary to control for any competing exposures that share a common ancestor with the exposure of interest. Hence, we believe that controlling for child diet is the appropriate statistical procedure to perform and this is a known method to extract direct effects, where intermediate variables account for some of the effect.

In the included articles that did not control for child diet, all but one (Barker et al, 2013) did not explicitly highlight that they were interested in the total effect of maternal and child diet. However, as mentioned in the manuscript, many of the articles are concerned with maternal intakes of fish but, with fish intake not being the primary exposure of concern (but reported to enough detail that it can be included in the meta-analysis), it might explain why child diet is accounted for

to such a small degree. (Though studies that investigate the direct effect of maternal fish-intake on child development should arguably also control for child diet.)

We have now amended the relevant section to better explain that; a) we are primarily interested in direct effects of maternal diet quality during pregnancy; b) the choice of control variables depends on the assumed causal model and c) we believe that a model in which maternal education is a shared predictor of maternal and child diet is most plausible, which means that one should control for child diet (which is a competing predictor). The text in the first paragraph under the heading "Maternal diet – direct effect or marker for child diet?" in the manuscript have been revised to the following: "An important issue that cannot be resolved by this meta-analysis is whether the observed association is based on direct effects of maternal diet quality during pregnancy or whether it is a marker for the child's diet, which is a competing exposure that also influences child development. Not surprisingly, maternal diet during pregnancy, as well as maternal post-natal diet, and child diet during infancy and early toddlerhood have been found to be highly correlated. Therefore, it remains possible that the observed associations between maternal diet quality during pregnancy and child development are due to the child's diet after pregnancy. Ultimately, the interpretation of the reported effect sizes depends on the assumed causal model. If it is assumed that child diet is a mediator between maternal diet quality during pregnancy and child development, then one has to control for child diet if interested in the direct effect of maternal diet quality during pregnancy, and one must not control for child diet if

interested in the total effect of maternal diet quality during pregnancy. However, we suggest that maternal diet and child diet have a common cause—e.g. parental education—and an unbiased estimate of the direct effects of maternal diet quality during pregnancy on child development requires controlling for child diet. One mitigating fact is that child diet varies with sociodemographic variables so that controlling for maternal postnatal diet and sociodemographic factors is likely to, at least in part, control for child diet. Still, child diet should ideally be assessed as a distinct factor."

10. Overall, I prefer the term "association" rather than "effect" in this observational meta-analysis.

ANSWER: We agree that using "association" might be more appropriate considering that "effect" might be perceived as indicating causal inference rather than an association. We have therefore taken greater care to be more precise in the use of the terms "effect", "effect size"and "association" as described in the last paragraph in the answer to question 7 above.

11. I am very surprised no effort was made to address study risk of bias. Publication bias is one form of potential bias, but the individual studies need to be assess for factors that might have led to biased results. I do not buy the authors' justification that because there is no "causal interpretation" attempted, quality issues of the included studies are irrelevant. All observational studies are limited for causal inference, but a systematic reviewer must give his/her readers an assessment of the quality of the included studies. This either needs to be completed, or a very strong rationale for not doing so needs to be presented in the paper itself.

ANSWER: Thank you for addressing this. We have now assessed the methodological quality of each included study with the Newcastle-Ottawa Scale (NOS) for assessing the quality of non-randomized studies in meta-analysis. We have included the scoring details for each study in supplementary table 3 and have added the following text under the new heading "Individual study quality assessment" in the methods section: "For each eligible study included in the meta-analysis we performed an individual study quality assessment using the NOS. The NOS provides an easy to use study quality checklist and is recognized by Cochrane. The scoring system is based on the assessment of three aspects of a study; Selection (representativeness of cohort and exposure assessment); Comparability (ascertainment of confounding); and Outcome (assessment of outcome and follow-up). The scoring system categorizes studies as being of good, fair or poor methodological quality, whereby insufficiency in one of the domains results in a "poor" rating. While the NOS has been criticized for an overly general definition of quality criteria, this generality allows for a wide application of the scale. Moreover, the intent of the scale is clear: A good rating of the Selection dimensions requires a representative sample and high quality measurement; a good rating of the Comparability dimension requires control of appropriate confounders; and a good rating of the Outcome dimensions requires a high quality measurement of outcomes and/or high follow up rates or correction for non-random drop out."

We have also added the following information in the results section, under the new heading "Individual study quality assessment" based on evaluation of the studies with NOS: "Each study was evaluated with the NOS checklist, please see supplementary table 3 for individual study scoring information. Of the 18 included studies, 9 were rated as of "fair" quality and 9 as of "poor" quality. No study received the rating "good" because none of the studies that used high quality measurements also adequately dealt with self-selection into studies and selective dropout of participants. All "poor" ratings were due to insufficiencies in the "Outcome" dimension of the NOS in studies that measured outcomes through self-reports (independent blind assessments or record linkage is preferred by the NOS) and that additionally did not account for selective dropout between exposure and outcome assessments."

12. Please confirm that in Figure 2, the ultimate diamond (0.13 [0.09, 0.17]) does not include results from the same study cohort, with the same measure in the same participants only at different times. For example, if Bernard et al 2013.1, Bernard et al 2013 each contained the same participants, then

you are double counting, and the overall effect estimate effectively double weights this one study (Bernard 2013). I do appreciate the efforts the authors have taken to attempt to explain this, but it is still not as clear as it should be to be confident in the forest plot.

ANSWER: Thank you for pointing this out. We have accounted for studies that report effect sizes in multiple dimension with appropriate analytical techniques, so that these studies are not counted more than once when calculating overall summary effect. We have done this by recalculating the variance and standard error for these studies based on a smaller N, before performing the meta-analytic model for the overall summary effect. For the studies that report one effect size in both the cognitive and affective domain, we have averaged the N across the two effect

sizes and then divided by two (((N1 +N2)/2)/2). For the studies with more than two reported effect sizes across domains, the denominators are equivalent to the total number of effect sizes included per study.

To make this explicit in the manuscript we have included the following text in the paragraph prior to figure 2: “For studies reporting more than one effect size, the appropriate statistical techniques were applied so that this was accounted for when calculating the summary effect size. When calculating the summary effect size in the two domains separately, a study that reported effect sizes for one cognitive and one affective outcome would have its originally calculated standard errors. When calculating the summary effect size across domains, a re-calculated standard error was used (the N across the two effect sizes were averaged and then divided by two (((N1 +N2)/2)/2), before calculating the adjusted variance and subsequently standard error. For the studies with more than two reported effect sizes across domains, the denominators were equivalent to the total number of effect sizes included per study). Hence, one study could be allocated different weights, depending on the analysis. In summary, whenever multiple effect sizes from one study entered the analysis, those were weighted correctly by adjusting the standard errors of the effect sizes accordingly.”

12. Was the search strategy developed with an information specialist/librarian? If so, please state this. If not, I suppose you can avoid mentioning it.

ANSWER: The search string was developed in conjunction with a specialist librarian. The first author developed the initial search string, which was then checked by the specialist librarian with regard to how well it worked for searches in the relevant databases and revised according to her feedback. To clarify this in the manuscript we have added the following text to the last paragraph under heading “Search criteria and strategies”: “The search string was developed by the first author in conjunction with a specialist librarian”

VERSION 3 – REVIEW

REVIEWER	AC Del Re Center for Innovation to Implementation VA Palo Alto Health Care System, USA
REVIEW RETURNED	17-Apr-2017

GENERAL COMMENTS	The authors meta-analytically examine the association between maternal diet quality and child cognitive/affective outcomes. This is a revised manuscript and, although not one of the original reviewers, I do have a couple of broad comments and suggestions that might help to improve the quality and clarity of the manuscript. Overall, I found this manuscript to be well written but it could benefit from some reorganization, along with tweaking the analyses, as I will detail below. 1. There are a fair amount of effect size (ES) dependencies that

	need to be accounted for. Dependencies are arising from at least two sources: multiple effect sizes per study and from 6 studies included that use the same sample. The authors report on a clever way to adjust the dependencies for multiple effect sizes but use no reference for the procedure. In fact, there are several established methods for accounting for dependency among ESs (Hunter JE, Schmidt FL. Methods of meta-analysis: Correcting error and bias in research findings. Sage Publications, Inc; 2004 and Gleser LJ, Olkin I. Stochastically dependent effect sizes. The handbook of research synthesis. 1994:339–355 and others, eg, borenstein, 2010) and one of these should be preferred. R statistical software has some packages that can aggregate studies based on these procedures. One other option would be to conduct separate univariate meta-analyses for each component, e.g., one meta-analysis for cognitive outcomes and a separate meta-analysis for affective outcomes (and combine the 3 affective factors for omnibus analysis and maybe test each level in a single moderator analysis with the 3 affective outcome levels). Regarding the other dependencies, either only one of the studies based on the same sample should be included or use a statistical correction based on robust standard errors (Hedges 2010), although the former might be preferred. 2. To improve statistical power for the omnibus and moderator analyses, the authors should consider restructuring the dataset and analyses. This would involve including all outcomes in an overall omnibus (which has been done) and then testing a few possible moderators, including “proxy” diet (proxy or not), outcome type (cognitive and affective), and possibly even child age of assessment (which could partially control for child diet quality). The information presented in Table 4 might better be served as exploratory analyses with the main focus on what is described above. This restructuring would still lead to ES dependencies, which could be handled by [from least preferred to best] (1) randomly selecting which outcome type to include when studies report both outcome types (not the best procedure because you lose information), (2) using Hedges robust SE procedure to correct of dependencies, or (3) conduct a multivariate meta-analysis, which will account for dependencies and draw upon the relationship between outcome types for each study and impute values for missingness. If these are considered, option 2 or 3 would be preferred to option 1. 3. In the computation of ESs section, I did appreciate the description of decisions about eliminating ES dependencies and applaud the authors for such detail. References Hedges LV, Tipton E, Johnson MC (2010). “Robust Variance Estimation in Meta-Regression With Dependent Effect Size Estimates.” Research Synthesis Methods, 1(1), 39–65.

VERSION 3 – AUTHOR RESPONSE

Reviewer Name: AC Del Re
Institution and Country: Center for Innovation to Implementation, VA Palo Alto Health Care System, USA Competing Interests: None declared

REVIEWER COMMENT 1. There are a fair amount of effect size (ES) dependencies that need to be accounted for. Dependencies are arising from at least two sources: multiple effect sizes per study and

from 6 studies included that use the same sample. The authors report on a clever way to adjust the dependencies for multiple effect sizes but use no reference for the procedure. In fact, there are several established methods for accounting for dependency among ESs (Hunter JE, Schmidt FL. Methods of meta-analysis: Correcting error and bias in research findings. Sage Publications, Inc; 2004 and Gleser LJ, Olkin I. Stochastically dependent effect sizes. The handbook of research synthesis. 1994:339–355 and others, eg, borenstein, 2010) and one of these should be preferred. R statistical software has some packages that can aggregate studies based on these procedures. One other option would be to conduct separate univariate meta-analyses for each component, e.g., one meta-analysis for cognitive outcomes and a separate meta-analysis for affective outcomes (and combine the 3 affective factors for omnibus analysis and maybe test each level in a single moderator analysis with the 3 affective outcome levels).
Regarding the other dependencies, either only one of the studies based on the same sample should be included or use a statistical correction based on robust standard errors (Hedges 2010), although the former might be preferred.

ANSWER: Thank you for pointing this out. As we understand from the reviewers comments, he propose that we need to perform two steps to account for the dependencies: use corrected weights and calculate robust standard errors. We had calculated the correct weights, and our approach of splitting the total N of a cluster into the different studies for calculation of effect size variance results in the same weights as equation (13) in the Hedges et al (2010) paper the reviewer refer to, if all studies have the same number of participants. When looking at results from longitudinal studies with different follow-up times (like most of the articles included in this meta-analysis) the number of participants will be lower for later follow-up times. Using equation (13) in Hedges et al. (2010) in these instances will lead to less efficient estimates, because studies with larger sample sizes would receive the same weight as studies with smaller sample size. We calculated weights such that articles that used larger sub-samples of a study received correspondingly larger weights. However, we do agree that we have failed to include a reference as a basis for this procedure. The text now refers to Hedges et al (2010) and explains that our weighting scheme is equivalent, with the exception that articles based on the same study can have different weights if these articles used sub-samples of different size from the study.
In addition, we had not calculated robust standard errors and we thank you for pointing this out. We have now performed the meta-analyses with the robust function in the Metafor package (which also refers to Hedges et al, 2010). The calculation of robust standard errors with this function did not lead to important changes in p-values, only slightly larger confidence intervals. Furthermore, the reviewer suggested that we calculate separate meta-analyses for the different domains in addition to an overall meta-analysis. This is precisely what we had done. That is, we calculated one meta-analysis for the cognitive domain, one for the affective domain, and one overall meta-analysis to estimate the overall summary effect size across all studies - each time using weights adjusted for the particular sample of included studies. It seems that the reviewer understood that the results in table 4 for “Affective”, “Cognitive”, and “summary effect size” come from one meta-regression, but these are in fact the results of the abovementioned three separate meta-analyses. However, after reviewing the manuscript we appreciate that we did not explain this explicitly enough in the text. We have addressed this in the revision by creating a new subtitle “Effect size dependencies” in the results section, where we have included the following text:
“There are several sources of effect size dependencies within our sample of studies. Firstly, some of the included studies use sub populations of the same cohort sample, while using different outcome measures at different time points: Four of the studies were based on subsamples of the ALSPAC cohort, and two studies were based on subsamples from the Project Viva cohort. Secondly, some of the studies included in this meta-analysis report multiple effect sizes. We used a weighting scheme and calculated robust standard errors to account for these sources of dependencies. Weights were adjusted for studies that contribute multiple effect sizes by recalculating them such that the sum of the weights of all effect sizes from a study reflect the sample size of that study. When using the Metafor

package, which calculates weights from effect size variances or standard errors, this can be achieved by calculating effect size variances with adjusted N. In particular, we adjusted N for study i such that $\tilde{N}_i = N_i / \sum_{j}^{k} N_j$. Here k is the number of effect sizes from a study sample (e.g. ALSPAC) and Nj are the sample sizes for the different effect sizes. When estimating average effect sizes for specific domains, this approach corrects for multiple effect sizes for one domain coming from one study sample. When estimating the overall effect size, this approach corrects for multiple contributions of effect sizes from one or more domains from one study sample. Hence, one study could be allocated different weights, depending on the meta-analytic model. Optimally the calculation of overall effect sizes would also account for the covariance between effects in different domains. However, the reviewed articles did not provide this information. The employed weighting scheme implies an assumed correlation of rho = 0.5. For the two REMs investigating for publication bias (trim and fill and with standard error as moderator) the adjusted standard errors based on the above formulae was used. For the original REM we used the reported effect sizes and corresponding adjusted variances as a basis for the calculations and obtained robust standard errors using Metafor's "robust" function. We chose this robust estimator function as it is appropriate to use for models with unspecified heteroscedasticity, which is the case with all studies reporting multiple effect sizes that are included in this meta-analysis."

REVIEWER COMMENT 2. To improve statistical power for the omnibus and moderator analyses, the authors should consider restructuring the dataset and analyses. This would involve including all outcomes in an overall omnibus (which has been done) and then testing a few possible moderators, including "proxy" diet (proxy or not), outcome type (cognitive and affective), and possibly even child age of assessment (which could partially control for child diet quality). The information presented in Table 4 might better be served as exploratory analyses with the main focus on what is described above. This restructuring would still lead to ES dependencies, which could be handled by [from least preferred to best] (1) randomly selecting which outcome type to include when studies report both outcome types (not the best procedure because you lose information), (2) using Hedges robust SE procedure to correct of dependencies, or (3) conduct a multivariate meta-analysis, which will account for dependencies and draw upon the relationship between outcome types for each study and impute values for missingness. If these are considered, option 2 or 3 would be preferred to option 1.

ANSWER: Thank you for pointing this out. Due to the limited amount of eligible studies for inclusion into the meta-analysis, the power issue is definitely something that has to be considered. As mentioned in the above reply to comment one, we had conducted three separate meta-analyses and we tested for several moderators, only in the meta-analyses that included all studies to yield the overall summary effect size. We initially considered three moderators (publication year, diet category (proxy or not) and instrument category (measurement of outcome - questionnaire or neuropsychological test) but ended up only using publication year and diet category as the heterogeneity was only present across the studies included in the affective domain, where all outcomes were measured by questionnaire. What we did not include is child age at assessment and outcome domain (cognitive or affective) as moderators in the overall analysis, but we have now included these moderators. We tested all four moderators in separate models, where only diet category was significant. However, we decided to include all moderators that explained some of the heterogeneity (R2) even though they were not significant, which were all but child age at assessment. Outcome domain, publication year and diet category explained a total of about 30 % of the heterogeneity present, and this overall moderator analysis was significant. We have included this updated information under the "moderator analyses" headline in the results section.

We believe that we have addressed the issue of power in a satisfactory manner, by using the reviewers proposed procedures for accounting for dependencies as mentioned in the above response to comment 1, which would align with the reviewers second suggestion in comment 2 performing three separate meta-analysis: one to yield an overall summary effect size, in addition to

calculating two separate meta-analyses for the affective domain and cognitive domain
After adding, as the reviewer suggested, child age at assessment and outcome domain as moderators, we believe that we have exhausted the different ways of handling the data. By explicitly presenting results from all the above-mentioned analysis, we believe that we have reported the results in a thorough and unambiguous matter. However, we do appreciate that we might not have explained this procedure in as much detail as we should have. We have therefore elaborated on this issue in the “limitations” by including the following text: “This meta-analysis used non-independent effect sizes due to both multiple effect sizes reported from the same study as well as several studies using the same cohort as a basis for their study sample. However, we took care in accounting for these dependencies by calculating adjusted weights and robust standard errors, and performing an overall meta-analysis as well as individual meta-analyses for both domains. Moreover, three different REMs were fit (c.f. Table 4) to test for plausible moderators. Hence, despite the challenges posed by dependent effect sizes our analyses likely provides unbiased summary effect sizes for the association between maternal diet quality and child development.”