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 effectiveness of evidence based treatments of Fetal Alcohol Spectrum Disorders in children and adolescents: A systematic review protocol
AUTHORS	Singal, Deepa; Menard, Chantalle; Neilson, Christine; Brownell, Marni; Hanlon-Dearman, Ana; Chudley, Albert; Zarychanski, Ryan; Abou-Setta, Ahmed

VERSION 1 – REVIEW

REVIEWER	Christie L. M. Petrenko, Ph.D. Mt. Hope Family Center, University of Rochester, USA
REVIEW RETURNED	12-Sep-2016

GENERAL COMMENTS	This manuscript clearly outlines a protocol for a systematic review of pharmacological and non-pharmacological interventions for FASD. Publishing protocols enhances the rigor of science and is a commendable step. The method of the proposed review appears generally sound. However, several issues reduce the enthusiasm for this proposed review. First, the growing body of literature to be reviewed is still quite limited, especially for pharmacological treatments, and the proposed analytical procedures (e.g., meta-analysis) may not be tenable or will produce limited results given the small number of studies and heterogeneity of measurement. Furthermore, the rationale for requiring a new systematic review when one was recently published (Reid et al., 2015) was not very convincing. Only a couple of intervention studies have been published since that review. The authors' primary argument was that Reid et al., did not include pharmacological interventions. However, only a couple of very small RCTs have been published on pharmacological interventions, which have been reviewed previously. A few other significant issues are worth noting for further consideration. First, the authors have selected measures of hyperactivity, impulsivity, and attention as primary outcomes. These measures may be representative of outcomes from the few stimulant pharmacological interventions, but are not commonly used for behavioral interventions for children with FASD. Most behavioral interventions are focused on teaching children and adults working with them new skills as well as making environmental accommodations to help compensate for primary disabilities.

	The examples listed for secondary outcomes measures (e.g., Personal Behavior Checklist, IQ, Ballard Tests, Depression Inventories) are not common across behavioral interventions and for the most part represent outcomes from a single study (Adnams et al., 2007). Measures used tend to be variable across studies and generally depend on the intervention targets (which vary considerably across studies). The most common measures tend to be things like the Eyberg Child Behavior Inventory (ECBI), Child Behavior Checklist (CBCL), Social Skills Rating System (SSRS), etc. Second, the types of side effects described for non-pharmacological trials are not the types of adverse outcomes that would be reported in behavioral trials generally. Finally, it appears interventions including nutritional supplementation (e.g., choline) don't meet inclusion criteria. Several rigorous trials have recently been conducted examining prenatal or postnatal choline supplementation on outcomes in infants and children. These interventions seem relevant to include. Other minor points: - Introduction, page 5: Fetal alcohol effects (FAE) and prenatal alcohol effects (PAE) are not recognized diagnoses in the continuum of FASD. FAE was previously used as a diagnosis, but is no longer used due to lack of specificity. - The term intellectual disability is now preferred over mental retardation. - Consider using the term "secondary conditions" instead of "secondary disabilities." While the term "secondary disabilities" was used in seminal studies on this topic, the term has gone out of favor in related fields, and the term "conditions" or "issues" is recommended. For example: The National Center on Health, Physical Activity, and Disability (http://www.ncpad.org/360/2050/Defining~Secondary~Conditions~for~People~with~Disabilities) addresses this in an article and uses the term "secondary conditions". - Other sources to consider for searching could include conference abstracts from the Research Society on Alcoholism, dating back likely from about 2005. NIH Reporter (https://projectreporter.nih.gov/reporter.cfm) could also indicate any current or recent intervention trials funded by NIH. Thank you for the opportunity to review this manuscript.

REVIEWER	Elizabeth Eggins The University of Queensland Australia
REVIEW RETURNED	31-Jan-2017

GENERAL COMMENTS	The systematic review protocol proposed by the authors is considered and well written. The introduction effectively contextualises the proposed review in terms of the importance of the topic and what is known about the area (including pre-existing reviews). In addition, by generally adhering to the PRISMA-P checklist and Cochrane Methodological Expectations of Cochrane Intervention Reviews (MECIR), the authors have composed quite a rigorous systematic review protocol.

	I have a few suggestions for the authors that will enhance the methodological rigour of the review protocol. I am mindful that BMJ may have length restrictions, so appreciate that not all suggestion can be incorporated to the level of depth that is seen in Cochrane or Campbell Collaboration Reviews. SEARCH DATE AND LANGUAGE At the time of writing the protocol, the search date (inception to July 2016) would have been suitable. Given we are now in 2017, the authors could move the search date to 31st December 2016. Will there be any language restrictions, or will the authors only include English articles? The discussion notes that the review will not have language bias, but there is no specification in the body of the protocol. SEARCH STRATEGY The authors have outlined quite a comprehensive search strategy for identifying studies. I would suggest, however, expanding the academic databases to a few other disciplinary specific/multidisciplinary databases to ensure studies from other disciplines are captured. For example: CINAHL, ProQuest Dissertation & Theses Global, Scopus, Project CORK, possibly some social work databases (e.g., Social Services Abstracts). Perhaps perusing some Cochrane or Campbell Collaboration protocols in the area of alcohol and drug research might be fruitful for identifying other sources. The search strategy provided for EMBASE was instructive for understanding the search approach. The following points may have already been considered by the authors, but this wasn't clear in the table. In general, I suggest the authors apply the PRESS Guidelines to their search strategy (McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., & Lefebvre, C. (2016). PRESS peer review of electronic search strategies: 2015 guideline statement. Journal of Clinical Epidemiology, 75, 40-46). Some of the points below are more pragmatic (e.g., points 3 and 4), but others are important in terms of ensuring the authors' search has appropriate sensitivity and does not fail to capture relevant research (e.g., points 2 and 5). 1. What is the search field for the first search line (MeSH or EMBASE equivalent or subject heading, will it be exploded)? 2. Use of wildcards to capture spelling variations, particularly for foetal and paediatric (some databases will automatically capture variations, whereas other do not). Important research may not be captured with different spelling 3. Use of wildcards to capture plural terms, instead of using multiple terms (e.g., "Foetal Alcohol Spectrum Disorder*" instead of "Foetal Alcohol Spectrum Disorder" and "Foetal Alcohol Spectrum Disorders". Generally databases will permit wildcards in phrases and many databases will automatically search for plurals (in case this was the reason for separate terms versus the one term). 4. The number of terms in line 4 could be reduced by search for a root word (unless there has been a specific reason for listing them separately). For example, child* could be used to capture children, childhood. In addition, there are some unnecessary duplications (e.g., minors). 5. Could some additional search terms around prenatal exposure to alcohol capture additional records? This could unnecessarily increase the sensitivity of the search, but could be tested by examining what additional research is captured that uses the new terms AND NOT the other search terms.

	INCLUSION CRITERIA/STUDY SELECTION Under the 'Study selection' section, the authors have quite clearly outlined how identified records and documents will be assessed for eligibility. However, it may be useful to explicate exactly how the eligibility criteria will be applied at each screening stage (i.e., titles/abstracts and then full-text documents). For example, will titles/abstracts be assessed against all criteria listed in Table 1 or a smaller set of criteria that are feasible to assess on title/abstract? Listed below are some other considerations in terms of the inclusion criteria according to PICO domains. Population: In terms of FASD populations, will the authors require an official diagnosis for study inclusion or will it be sufficient if study authors state that the study is using an FASD sample? Research Design: The authors correctly identify in their Discussion that RCTs are the gold standard for determining treatment effectiveness. However, I do wonder, given the area, whether some more rigorous quasi-experimental designs should be included so that practitioners and other researchers gain an understanding of the breadth and, possibly, the limitations of the existing evaluation evidence. Intervention: Will any type of intervention setting be eligible for inclusion (e.g., school, home, clinic etc?) Outcomes: The authors state that outcome measures need to be 'validated'. I was curious as to how this will be operationalised in the review when making inclusion/exclusion decisions. For example, would education outcomes (secondary outcome) measured using official education data be considered 'validated'? On a related note, I am curious why only standardised cognitive outcomes will be included? Perhaps some cognitive measures with sound psychometric properties would be useful and provide insight into the types of outcomes interventions could address. ANALYSIS A little more depth is required for the analysis sections, particularly in the following areas: 1. Unit of analysis issues and how they will be handled (e.g., multiple eligible outcome measures, multiple reports for a study, clustered designs, multiple time points) 2. Consideration of how different intervention approaches and outcomes will be synthesised. For example, will pharmacological and non-pharmacological interventions be meta-analysed separately or together? There is the potential for this review to include a large range of heterogeneous interventions and outcomes. Given this, I would suggest the authors give some thought to how this will be handled in terms of the analysis (e.g., it may not be appropriate to combine some interventions or outcomes, necessitating the need for separate meta-analyses or more advanced approaches like network meta-analysis). 3. On its own, I2 is not sufficient for assessment of heterogeneity. I suggest the authors consult the Cochrane Handbook and explore additional metrics (e.g., Tau and Chi2). Two final notes: the authors checked in PRISMA-P that they had specified how they will assess the quality of the evidence (p. 15?) and the type synthesis that will be conducted if meta-analysis is not appropriate (p. 14?). However, I could not locate this in the manuscript. There is content in the first paragraph of the discussion noting how RCTs are high quality evidence, but assessing the quality of the evidence expands on this and is different from assessing the risk of bias in individual studies. I suggest the authors examine GRADE on this point.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Christie L. M. Petrenko

Comment: This manuscript clearly outlines a protocol for a systematic review of pharmacological and nonpharmacological interventions for FASD. Publishing protocols enhances the rigor of science and is a commendable step. The method of the proposed review appears generally sound.

Response: Thank you for your thoughtful comments.

Comment: proposed review. First, the growing body of literature to be reviewed is still quite limited, especially for pharmacological treatments, and the proposed analytical procedures (e.g., meta-analysis) may not be tenable or will produce limited results given the small number of studies and heterogeneity of measurement. Furthermore, the rationale for requiring a new systematic review when one was recently published (Reid et al., 2015) was not very convincing. Only a couple of intervention studies have been published since that review. The authors' primary argument was that Reid et al., did not include pharmacological interventions. However, only a couple of very small RCTs have been published on pharmacological interventions, which have been reviewed previously.

Response: We agree with you that the body of literature is quite limited and do not expect to be able to conduct meta-analysis. We will not conduct meta-analysis if the studies are heterogeneous; will generate a qualitative appraisal and review of results. We appreciate a similar review was published in 2015, however, our review was initiated before this review (as evident by the registration of our protocol on PROSPERO in 2013: CRD42013005996). The authors of the 2015 review did not have a published protocol so we could not see that they were conducting duplicate work. Due to the hard work put in by our study team we are still enthusiastic that our review will offer a new contribution to the literature by including pharmacological interventions, which are indeed the gold standard in medical literature. We are also including in our discussion section, a minimal scientific criteria to guide the next generation of trials in this area to provide more valid and clinically useful information, which is a novel contribution by our review.

Comment: A few other significant issues are worth noting for further consideration. First, the authors have selected measures of hyperactivity, impulsivity, and attention as primary outcomes. These measures may be representative of outcomes from the few stimulant pharmacological interventions, but are not commonly used for behavioural interventions for children with FASD. Most behavioural interventions are focused on teaching children and adults working with them new skills as well as making environmental accommodations to help compensate for primary disabilities. The examples listed for secondary outcomes measures (e.g., Personal Behavior Checklist, IQ, Ballard Tests, Depression Inventories) are not common across behavioral interventions and for the most part represent outcomes from a single study (Adnams et al., 2007). Measures used tend to be variable across studies and generally depend on the intervention targets (which vary considerably across studies). The most common measures tend to be things like the Eyberg Child Behavior Inventory (ECBI), Child Behavior Checklist (CBCL), Social Skills Rating System (SSRS), etc.

Response: We do agree with you that measures are extremely variable across studies and have not intended on our list to be exclusive or limited to the scales we have listed. These were intended to provide examples and we will include all other measures and outcomes that studies include. We have clarified this in the manuscript, and added your suggestions for the most common measures.
The following paragraph has been added: "It is important to note that measures used in this field of study tend to be variable and depend on the intervention targets.

We have provided examples of measures that may be included in studies, however, we do not intended this list to be exclusive and will include trials with any standardized measures, including Eyeberg Child Behaviour Inventory (ECBI), Child Behaviour Checklist (CBCL), Social Skills Rating System (SSRS)."

Comment: Second, the types of side effects described for nonpharmacological trials are not the types of adverse outcomes that would be reported in behavioral trials generally.

Response: We have expanded this section to include the following adverse outcomes that have been noted in the literature as common side effects from behavioral/psychological trials. "Side effects of non-pharmacological treatments including, but not limited to: increase in symptoms or development of new symptoms caused by the behavioral/psychological intervention including: increase in psychiatric symptoms, increase in ADHD symptoms, increase in agitation and decrease in social skills."

Comment: Finally, it appears interventions including nutritional supplementation (e.g., choline) don't meet inclusion criteria. Several rigorous trials have recently been conducted examining prenatal or postnatal choline supplementation on outcomes in infants and children. These interventions seem relevant to include.

Response: Thank you for pointing out this oversight, we will include nutritional supplements as part of our non-pharmacological interventions and include trials that have looked at choline, as this is an important area that is also not addressed in other reviews. Including trials that have looked at this intervention would also make a novel contribution to this review. However,
we do want to only examine interventions in children diagnosed with FASD, therefore, if these trials were conducted on children exposed to prenatal alcohol but did not have an FASD diagnosis we cannot include them as we want our recommendation to be clinically relevant to patients with a diagnosis of FASD. We have included the following: "Trials evaluating the effect of nutritional supplements (such as choline) will also be included".

Comment: Other minor points:
Introduction, page 5: Fetal alcohol effects (FAE) and prenatal alcohol effects (PAE) are not recognized diagnoses in the continuum of FASD. FAE was previously used as a diagnosis, but is no longer used due to lack of specificity.

Response: Thank you for pointing out this important oversight; we have updated our manuscript using the new updated Canadian guidelines by Cook et al, 2015. The following changes have been made: "Prenatal alcohol use places children at risk for Fetal alcohol Spectrum Disorder, an umbrella term…." Umbrella term has been changed to "diagnostic term" as per the updated guidelines and the new guidelines have been added to the reference list.

Comment: The term intellectual disability is now preferred over
mental retardation.

Response: We have changed mental retardation to intellectual disability.

Comment: Consider using the term "secondary conditions" instead of "secondary disabilities." While the term "secondary disabilities" was used in seminal studies on this topic, the term has gone out of favor in related fields, and the term "conditions" or "issues" is recommended. For example: The National Center on Health, Physical Activity, and Disability (http://www.ncpad.org/360/2050/Defining~Secondary~Co nditions~for~People~with~Disabilities) addresses this in an article and uses the term "secondary conditions".

Response: We have edited the paper to use the term “secondary conditions” instead of secondary disabilities.

Comment: Other sources to consider for searching could include conference abstracts from the Research Society on Alcoholism, dating back likely from about 2005. NIH Reporter (https://projectreporter.nih.gov/reporter.cfm) could also indicate any current or recent intervention trials funded by NIH.

Response: We have included this in our search.

Comment: Other minor points: Introduction, page 5: Fetal alcohol effects (FAE) and prenatal alcohol effects (PAE) are not recognized diagnoses in the continuum of FASD. FAE was previously used as a diagnosis, but is no longer used due to lack of specificity.

Response: Thank you for pointing out this important oversight; we have updated our manuscript using the new updated Canadian guidelines by Cook et al, 2015. The following changes have been made: “Prenatal alcohol use places children at risk for Fetal alcohol Spectrum Disorder, an umbrella term….” Umbrella term has been changed to “diagnostic term” as per the updated guidelines and the new guidelines have been added to the reference list.

Comment: The term intellectual disability is now preferred over mental retardation

Response: We have changed mental retardation to intellectual disability.

Comment: Consider using the term “secondary conditions” instead of “secondary disabilities.” While the term “secondary disabilities” was used in seminal studies on this topic, the term has gone out of favor in related fields, and the term “conditions” or “issues” is recommended. For example: The National Center on Health, Physical Activity, and Disability (http://www.ncpad.org/360/2050/Defining~Secondary~Co nditions~for~People~with~Disabilities) addresses this

Response: We have edited the paper to use the term “secondary conditions” instead of secondary disabilities.

Comment: Other sources to consider for searching could include conference abstracts from the Research Society on Alcoholism, dating back likely from about 2005. NIH Reporter (https://projectreporter.nih.gov/reporter.cfm)

Response: We have included this in our search

Reviewer 2:
Comment:
SEARCH DATE AND LANGUAGE
At the time of writing the protocol, the search date (inception to July 2016) would have been suitable. Given we are now in 2017, the authors could move the search date to 31st December 2016. Will there be any language restrictions, or will the authors only include English articles? The discussion notes that the review will not have language bias, but there is no specification in the body of the protocol.

Response: We have changed the date to March 2017. There will be no language restrictions, we have added this to the manuscript.

Comment: SEARCH STRATEGY
The authors have outlined quite a comprehensive search strategy for identifying studies. I would suggest, however, expanding the academic databases to a few other disciplinary specific/multidisciplinary databases to ensure studies from other disciplines are captured. For example: CINAHL, ProQuest Dissertation & Theses Global, Scopus, Project CORK, possibly some social work databases (e.g., Social Services Abstracts). Perhaps perusing some Cochrane or Campbell Collaboration protocols in the area of alcohol and drug research might be fruitful for identifying other sources.

Response: Thank you for your helpful suggestions. We have modified our search strategy to incorporate your suggestions. We will utilize: MEDLINE (Ovid), CINHAL Plus with Fulltext (EBSCO), CENTRAL (Cochrane Library – Wiley), PsycINFO (ProQuest), and Proquest Dissertations & Theses (Proquest).

Comment: The search strategy provided for EMBASE was instructive for understanding the search approach. The following points may have already been considered by the authors, but this wasn't clear in the table. In general, I suggest the authors apply the PRESS Guidelines to their search strategy. Some of the points below are more pragmatic (e.g., points 3 and 4), but others are important in terms of ensuring the authors' search has appropriate sensitivity and does not fail to capture relevant research (e.g., points 2 and5).

Response: See Table 2.

Comment: 1. What is the search field for the first search line (MeSH or EMBASE equivalent or subject heading, will it be exploded)?

Response: See Table 2.

Comment: 2. Use of wildcards to capture spelling variations, particularly for foetal and paediatric (some databases will automatically capture variations, whereas other do not). Important research may not be captured with different Spelling

Response: Have incorporated the use of wildcards. Table 2

Comment: 3. Use of wildcards to capture plural terms, instead of using multiple terms (e.g., "Foetal Alcohol Spectrum Disorder*" instead of "Foetal Alcohol Spectrum Disorder" and "Foetal Alcohol Spectrum Disorders". Generally databases will permit wildcards in phrases and many databases will automatically search for plurals (in case this was the reason for separate terms versus the one term).

Response: Have incorporated the use of wildcards. Table 2

Comment: 4. The number of terms in line 4 could be reduced by search for a root word (unless there has been a specific reason for listing them separately). For example, child* could be used to capture children, childhood. In addition, there are some unnecessary duplications (e.g., minors).

Response: Have incorporated. See attached.

Comment: 5.Could some additional search terms around prenatal exposure to alcohol capture additional records? This could unnecessarily increase the sensitivity of the search, but could be tested by examining what additional research is captured that uses the new terms AND NOT the other search terms.

Response: Thank you for this suggestion, we will consider this sensitivity analysis when revising our final search with our librarian.

Comment: INCLUSION CRITERIA/STUDY SELECTION
Under the 'Study selection' section, the authors have quite clearly outlined how identified records and documents will be assessed for eligibility. However, it may be useful to explicate exactly how the eligibility criteria will be applied at each screening stage (i.e., titles/abstracts and then full-text documents). For example, will titles/abstracts be assessed against all criteria listed in Table 1 or a smaller set of criteria that are feasible toassess on title/abstract?

Response: We have added the following statement to clarify our review process: "The reviewers will assess titles/abstracts against a smaller set of criteria that are more feasible to assess title/abstract".

Comment: Listed below are some other considerations in terms of the inclusion criteria according to PICO domains. Population: In terms of FASD populations, will the authors require an official diagnosis for study inclusion or will it be sufficient if study authors state that the study is using an FASD sample?

Response: We have included the following clarification in our population statement: "Children (<18 years), both males and females, with an official diagnosis (i.e from a physician) of FASD including"

Comment: Research Design: The authors correctly identify in their Discussion that RCTs are the gold standard for determining treatment effectiveness. However, I do wonder, given the area, whether some more rigorous quasi-experimental designs should be included so that practitioners and other researchers gain an understanding of the breadth and, possibly, the limitations of the existing evaluation evidence.

Response: While we agree that rigorous quasi-experimental designs are of importance in this area, this is beyond the scope of this review as we would like to highlight this significant gap for the international readers and stress the importance that future trials that need to be conducted in this area in order to move this field forward. We do agree of the importance of rigorous qausiexperimental designs in this area and we are conducting future work that will include non-RCT studies.

Comment: Intervention: Will any type of intervention setting be eligible for inclusion (e.g., school, home, clinic etc?)

Response: Yes

Comment: Outcomes: The authors state that outcome measures need to be 'validated'. I was curious as to how this will be operationalised in the review when making inclusion/exclusion decisions. For example, would education outcomes (secondary outcome) measured using official education data be considered 'validated'? On a related note, I am curious why only standardised cognitive outcomes will be included? Perhaps some cognitive measures with sound psychometric properties would be useful and provide insight into the types of outcomes interventions could address

Response: We agree that we cannot ensure that all outcome measures studies utilize will be validated, and we do not want to include studies that have non-validated scales or measures, therefore we have taken out the word "validated" from this statement. We will comment on whether the scales were validated or not in our assessment of the quality of evidence produced by the studies.

Comment: ANALYSIS
A little more depth is required for the analysis sections, particularly in the following areas:

1. Unit of analysis issues and how they will be handled (e.g., multiple eligible outcome measures, multiple reports for a study, clustered designs, multiple time points)

Response: We completely agree that possible unit-of-analysis error should be anticipated and as such, we propose the following in accordance to the Cochrane Handbook: (we have added these statements to the data analysis section):
• For all studies, we will only include data from the longest follow-up period reported;
• For cluster-randomized trials, we will adjust the reported outcomes to account for the clustering using the interclass correlation coefficient. If this is not reported, then we will use a range of plausible ICCs and conduct sensitivity analyses to test the roubustness of the reported analyses; • For cross-over studies, we will adjust for the lack of independence of the units (similar to the clusterrandomized)

Comment: 2. Consideration of how different intervention approaches and outcomes will be synthesized. For example, will pharmacological and non-pharmacological interventions be meta-analysed separately or together? There is the potential for this review to include a large range of heterogeneous interventions and outcomes. Given this, I would suggest the authors give some thought to how this will be handled in terms of the analysis (e.g., it may not be appropriate to combine some interventions or outcomes, necessitating the need for separate meta-analyses or more advanced approaches like network meta-analysis).

Response: We intend on analyzing data from pharmacological and nonpharmacological interventions separately. Within each analysis interventions will be grouped based on relevant categories. At this time, we do not believe that the evidence is strong enough to require a network meta-analysis.

Comment: 3. On its own, I2 is not sufficient for assessment of heterogeneity. I suggest the authors consult the Cochrane Handbook and explore additional metrics (e.g., Tau and Chi2).

Response: We will review both the clinical and statistical heterogeneity (based on Chi-squared, Tau-squared and the I-squared statistic). For I-squared, we will also review the uncertainty intervals. We edited this section to state: “We will review both the clinical and statistical heterogeneity of the data using the Chi-squared Tau-squared and the I-squared statistics. For the I-squared test, we will also review the uncertainty intervals. If significant heterogeneity is suspected, further analysis including subgroup analysis will be conducted.”

Comment: Two final notes: the authors checked in PRISMA-P that they had specified how they will assess the quality of the evidence (p. 15?) and the type synthesis that will be conducted if meta-analysis is not appropriate (p. 14?). However, I could not locate this in the manuscript. There is content in the first paragraph of the discussion noting how RCTs are high quality evidence, but assessing the quality of the evidence expands on this and is different from assessing the risk of bias in individual studies. I suggest the authors examine GRADE on this point.

Response: We will not be conducting quality of the evidence.

VERSION 2 – REVIEW

REVIEWER	Christie Petrenko Mt. Hope Family Center University of Rochester USA
REVIEW RETURNED	01-May-2017

GENERAL COMMENTS	The authors attempted to address this and the other reviewers' comments, with some satisfactory edits. However, a few comments seem insufficiently addressed and there were several inconsistencies identified in the paper. Additional comments are also offered based on newly added details. In the introduction, the authors cite the newly revised 2015 Canadian guidelines when defining FASD and then refer to terms on the fetal alcohol spectrum that are used in the Institute of Medicine/Hoyme guidelines. The authors may want to consider a more nuanced discussion of FASD diagnosis. There are also other diagnostic terms under the umbrella of FASD reported in the literature (e.g., static encephalopathy, alcohol exposed; neurobehavioral disorder, alcohol exposure). A large number of citations are provided for prevalence estimates in the first paragraph. It is recommended to put the May et al., 2014 [4] citation with the other citations versus having it separate. There are also several citations listed in [5-14] that do not appear to empirically assess prevalence. The term mental retardation is still used in line 35 on page 5. Intellectual disability is the preferred term. What is the rationale for including children only with a formally diagnosed FASD? At first glance this seems appropriate; however, diagnosis of young children (especially under 3) is often not possible until they are older and cognitive and behavioral impairments emerge. Only including children with a formal FASD diagnosis could potentially limit inclusion of studies targeting this key developmental period. Similarly, will interventions delivered during the prenatal period be included when infant/child outcomes are reported? For example, prenatal choline supplementation has been rigorously tested and infant outcomes have been examined. Will studies that compare two active interventions (e.g., PCIT and Parent psychoeducation/support groups) be included? What is the rationale for focusing on inattention, hyperactivity, and impulsivity as primary outcomes? As mentioned previously, these are not the primary outcomes of most behavioral interventions. In addition, they represent only a small fraction of the types of primary disabilities present in children with FASD. Some would argue that these are not necessarily the most relevant symptoms experienced by children with FASD that impact quality of life for children and families. A stronger rationale is needed for the primary focus on these specific symptoms or the authors may consider broadening the list of primary outcomes. Also, authors indicate primary measures will rely entirely on rating scales (vs. neuropsychological child measures or other objective assessments). This seems limited. In contrast to pharmacological interventions, the adverse outcomes identified for non-pharmacological interventions do not appear to come from separate measures/outcomes than targeted primary/secondary outcomes. But rather, they generally reflect the opposite direction of change that what is targeted. Things like child maltreatment, placement changes, or suicidal ideation/attempts

	could be considered. What is the rationale for only reporting the longest follow-up period? It would seem of interest to consider both the immediate post-intervention and longest follow-up period to assess the initial strength of effects and whether effects changed over time (continued to improve, maintain, dissipate). This is particularly relevant for behavioral interventions where behavior change agents (i.e., often parent or teacher) continue to implement (or not) targeted strategies. Also, what will be done in the common situation when there is a delayed-waitlist comparison group who receives the intervention in the interval between immediate post-intervention and subsequent follow-up? The authors indicate they plan to search conference proceedings from the last 3 years for unpublished results. This reviewer is aware of multiple trials that had strong designs, but have yet to be published dating back to the mid- to late 2000 for various reasons. I'd recommend considering going back further. The most relevant conferences would be the Research Society on Alcoholism (abstracts published in Alcoholism: Clinical and Experimental Research) and the International Conference on FASD hosted by the University of British Columbia every other year (2017 was 7th conference). On page 12, the authors added that a smaller set of criteria will be used to initially assess inclusion based on the abstract and title. This seems appropriate; however, this sentence is vague. What criteria will be used? The second paragraph on page 12 includes content about the form being piloted that is repetitive from the prior paragraph. On page 14, the authors state they will contact study authors for missing data. Do they mean missing data from participants (e.g., due to attrition), unpublished data on other outcomes, or statistics not reported in the paper needed for effect size calculations? They also stated that they will solicit unpublished data from authors in progress or recently completed trials. It seems unlikely researchers will feel comfortable sharing in progress or recently completed data before they have had the chance to present or publish on the data themselves first. In response to reviewer 2's final question, the authors stated they will not be reviewing the quality of the evidence. However, in a prior comment they stated they will be assessing the quality of the evidence when commenting on the validity of measurement scales. The first sentence in the discussion also alludes to using a standardized rating tool to assess methodological quality of studies (assuming this is referring to the bias tool?). This is inconsistent. The authors have decided to only include studies using an RCT design, which omits some rigorous and well-done studies (e.g., O'Connor et al., 2006 – used alternating assignment within cohorts). Most "real-world" effectiveness in community settings would also be excluded with this criterion, which some may argue is even more important than RCTs in more controlled laboratory settings. In addition, the authors state they intend to put forth "minimal scientific criteria to guide the next generation of trials in this area to provide

	more valid and clinically useful information." This seems to imply that the current body of work is inadequate (despite being small) and that studies using alternate designs are not worthy of consideration. While criteria could be theoretically useful, this reviewer has some concerns that such criteria could hamper future growth of intervention research rather than promote the next generation. Conducting RCTs of interventions for FASD should be encouraged and promoted, but there are limitations to this design in some regards and practical challenging realities (e.g., funding, access to large sample sizes). Progress on empirically validated interventions for people with FASD was almost nonexistent for decades and has only shown gradual growth over the last 10 years. If the authors decide to advocate for minimal scientific criteria, they may want to consider a diversity of designs that promote a logical and systematic approach to intervention development and evaluation. Olson et al., 2016 (ACER) may be useful in considering key research questions for intervention research and potential designs to address these key questions. A diversity of acceptable designs also provides greater opportunities for a broader range of researchers and clinicians to contribute to the scientific enterprise. Finally, there are several locations where a word or punctuation is missing. A thorough edit is recommended.

REVIEWER	Elizabeth Eggins University of Queensland, Australia.
REVIEW RETURNED	06-May-2017

GENERAL COMMENTS	SEARCH STRATEGY The authors have addressed all areas raised in the previous peer-review. The use of forward citation searching for 'key' studies is somewhat vague and could introduce bias if the operationalisation of 'key studies' is not made explicit in the protocol. I suggest the authors conduct forward citation searching for all studies deemed eligible for review. • Thank you for your comments. We agree that our current statement is vague and have edited it to state that we will be conducting a forward citation search for all studies that are eligible for review. "We will also perform forward searches of all studies included in this review in Web of Science to identify additional citations that might have been missed in the database search" page 11 INCLUSION CRITERIA/STUDY SELECTION 1. Title/Abstract Screening. The authors have identified that a reduced number of criteria will be used to screen titles/abstracts. Could the authors provide some more explicit detail for what criteria will be used? This may require some piloting. It can be difficult to unequivocally decide on inclusion/exclusion based on titles and abstracts, so it would be helpful to understand exactly how the authors will conduct this stage of their review in order to assess whether their decided method is sound. • We agree that our statement "we will use a reduced number of criteria" is vague, therefore we have taken steps to clarify and provide more detail. It is common in systematic reviews to only look at the population, intervention and study designs during the title/abstract screening phase. We have added that we will be looking at the population, intervention and study design during this

	phase to provide more detail for our readers and removed the statement “a smaller set of criteria” “The reviewers will assess titles/abstracts for studies that meet criteria for: population, intervention and study design”. Page 12 2. Intervention Setting. If all types of intervention settings are to be included in the review, the authors need to explicitly state this in the protocol. • We have included the following statement on page 8 “Types of Interventions: All pharmacological or non-pharmacological interventions targeting the improvement of FASD symptoms in children in all types of intervention settings will be included” 3. FASD Diagnosis. Requiring a formal FASD diagnosis could be problematic and unnecessarily exclude some research. Some authors may not explicitly report whether the diagnosis is official or if there even has been a diagnosis (e.g., authors may just state they are using a cohort of children with FASD without any detail on diagnosis). I would suggest including studies that state they are using an FASD sample and studies that report directly on diagnosis. The authors could then code whether or not there is an official diagnosis and see if results differ during the analysis phase. • Thank you for your helpful suggestions. We agree with you that obtaining data on whether the diagnosis of FASD is “official” is difficult. We have taken your suggestion and have edited the following statement on page 8 to read: “Types of Participants: Children (<18 years), both males and females, with an author defined diagnosis of FASD (i.e studies which state they are using a cohort of children with FASD) including but not limited to: Fetal Alcohol Syndrome, Partial Fetal Alcohol Syndrome, Alcohol-related Neurodevelopmental Disorder, and Alcohol-related Birth Defects.” 4. Outcomes. It is till not clear whether validated and unvalidated measures are included. The authors’ response to this point is not clear and some of the 'validated' terminology has been removed (e.g., in the abstract), but in other areas it has not (e.g., outcomes section). Moreover, it is still not clear by what the authors mean by 'validated'. How exactly will this be operationalised when making inclusion/exclusion criteria? Is reliability also important? What type of validity? If this data is not reported in the study, will the study be excluded or will the authors verify psychometric qualities by looking at broader literature for the measure? The authors have also not sufficiently addressed the point on standardised cognitive outcomes. To provide a comprehensive review, I would suggest including both standardised/non- standardised and validated/unvalidated measures. As the authors correctly state, the quality of the measures can be evaluated during the risk of bias assessment. Moreover, the type/quality of outcome measures could be used in a subgroup analysis to examine whether quality of outcome measures impacts the overall impact of interventions. • Thank you for your helpful suggestions. We will evaluate all outcomes presented in all included studies, we are not excluding studies or outcomes based on the use of validated or unvalidated/

	standaridized or non-standardized measures. • We have stated in the outcome measures section we will include studies that are pertaining to the children's physical and mental health, as well as cognitive, behavioural and social skills which are presented in included studies. These outcomes may be measured using standardized/non-standardized and validated/un-validated measures, for example by rating scales (e.g. Child Behaviour checklist). All outcome measures included in the studies will be reported in this review. We have added this to the outcomes section and have removed any specific statements to only validated or standardized measures. ANALYSIS AND SYNTHESIS 1. Unit of analysis. I am curious as to why the authors would only include the longest follow-up time-point. It would be useful to examine whether effects are seen at post-intervention and then maintain at follow-up time-points. This can be achieved by conducting separate meta- analyses at specific time-points. Perhaps the authors could examine some recent Cochrane reviews that have taken this approach. Examining the impact of interventions over time is an important practice and policy issue. Practitioners and policy-makers need to understand if effects take some time to appear, appear immediately, reduce over time, or improve over time. • We agree with the reviewers comments and have edited the sentence on page 13 to read: "For all studies, we will include data from all reported time periods" 2. Coding of intervention types for analysis. Could the authors please provide their General approach for how non-pharmacological interventions will be grouped? This could be quite general and informed by previous reviews, but does require some attention so that it can be assessed as part of the overall review protocol (e.g., categorisation by modality or anticipated theoretical underpinning/approach). • Page 8 "Non-pharmacological interventions will be grouped according to categorization by type of intervention, i.e behavioral intervention, educational intervention, social intervention". 3. The authors have not addressed the first point on how they will synthesise the evidence if meta-analysis is not appropriate. Moreover, I am curious as to why the authors will not assessing the overall quality of the evidence when this is now considered an important component of systematic reviews. Could the authors please provide a justification for this or include their plans for conducting an appraisal of the body of evidence (latter more acceptable approach). • If meta-analysis is not appropriate, we will conduct a quailative/narrative synthesis. As for narrative synthesis, there is limited guidance or standardization outside the conduct of a formal qualitative synthesis. It has been reported that while narrative synthesis are used in as much as half or more of systematic reviews in certain areas, their reporting is often inadequate (Campbell, Lancet, 2016). We will use guidance provided by several organizations in the conduct/ reporting of a narrative summary (if needed). Examples of the guidance documents including Cochrane (Ryan R; Cochrane Consumers and Communication Review Group. 'Cochrane Consumers and Communication Review Group: data synthesis and analysis'. http://cccrg.cochrane.org, June 2013), Lancaster University (Popay et al, Guidance on the Conduct of Narrative Synthesis in Systematic

	Reviews: A Product from the ESRC Methods Programme, 2006) and the NIHR Complex Reviews Support Unit (http://www.nihrcrsu.org/guidance/narrative_synthesis/). • We are going to evaluate the strength of the overall evidence using GRADE and have include this in our protocol on page 14: "Grading the Evidence for Each Primary Outcome: The strength of evidence for the primary outcomes will be graded by using the approach described by the GRADE working group. Two reviewers will evaluate the strength of a body of evidence independently, and discrepancies will be resolved through consensus. This approach assesses the evidence based on four domains: risk of bias, inconsistency, indirectness, imprecision, publication bias, and other factors (and upgrading). We will classify the strength of evidence as "high", "moderate", "low", or "very low" and make recommendations for future research needs.

VERSION 2 – AUTHOR RESPONSE

Reviewer 1:

In the introduction, the authors cite the newly revised 2015 Canadian guidelines when defining FASD and then refer to terms on the fetal alcohol spectrum that are used in the Institute of Medicine/Hoyme guidelines. The authors may want to consider a more nuanced discussion of FASD diagnosis. There are also other diagnostic terms under the umbrella of FASD reported in the literature (e.g static encephalophaty, alcohol exposed, neurobehavrioual disorder, alcohol exposure).
• Thank you for your comments. We have used information from the Canadian FASD Guidelines and have edited our paper to read (page 5)
"Diagnosis requires a neurodevelopmental assessment conducted by a multidisciplinary team and includes a social and medical history, along with complete physical examination. Patients with a diagnosis of FASD must have the confirmation of prenatal alcohol exposure and may have sentinel facial features and/or evidence of impairment in neurodevelopmental domains. More information regarding diagnosing FASD is available in the updated Canadian FASD guidelines".

A large number of citations are provided for prevalence estimates in the first paragraph. It is recommended to put the May et al., 2014 [4] citation with the other citations versus having it separate. There are also several citations listed in [5-14] that do not appear to empirically assess prevalence.
• Thank you for bringing this to our attention. We have re-written this portion of the introduction and made it more concise and clear by using a recent reference: "Prevalence of alcohol consumption during pregnancy and Fetal Alcohol Spectrum Disorders among the general and Aboriginal populations in Canada and the United States" by Poopova et al 2017, we have edited the previous sentence to now read:
• "FASD has been estimated at 5 in 1,000 people in Canada, and 15 in 1,000 people in the United States" – page 5

The term Mental retardation is still used in line 35 on page 5, Intellectual disability is the preferred term.
• We have removed the term and replaced it with intellectual disability – page 5

What is the rationale for including children only with a formally diagnosed FASD? At first glance this seems appropriate; however, diagnosis of young children (especially under 3) is often not possible until they are older and cognitive and behavioral impairments emerge. Only including children with a

formal FASD diagnosis could potentially limit inclusion of studies targeting this key developmental period.
• We do see the reviewers important point about missing studies that may be evaluating treatments on children without a formal diagnosis of FASD, but who exhibit symptoms, therefore we have edited our paper to include studies that evaluate all treatments for FASD on a “author defined” definition of FASD (see page 8).

Similarly, will interventions delivered during the prenatal period be included when infant/child outcomes are reported? For example, prenatal choline supplementation has been rigorously tested and infant outcomes have been examined.
• Although the effectiveness of prenatal interventions is important, it is outside the scope of this review, as we are not planning to include prophylactic interventions during the prenatal period.

Will studies that compare two active interventions (e.g., PCIT and Parent psychoeducation/support groups) be included?

• Yes, we will include all interventions presented in this patient population that are assessed in the studies included in our review.

What is the rationale for focusing on inattention, hyperactivity, and impulsivity as primary outcomes? As mentioned previously, these are not the primary outcomes of most behavioral interventions. In addition, they represent only a small fraction of the types of primary disabilities present in children with FASD. Some would argue that these are not necessarily the most relevant symptoms experienced by children with FASD that impact quality of life for children and families. A stronger rationale is needed for the primary focus on these specific symptoms or the authors may consider broadening the list of primary outcomes. Also, authors indicate primary measures will rely entirely on rating scales (vs. neuropsychological child measures or other objective assessments). This seems limited.
• We are interested in all outcomes related to this population and will include all outcomes that are presented in the included reviews. We will also review all primary measures, not limited to only rating scales. We have clarified this throughout the manuscript as well as changed the primary outcomes in this protocol to be more reflective of the disabilities and most relevant symptoms experienced by children with FASD. It is important to note that these outcomes listed are meant to be examples and not exclusive. The following changes have been made to the outcomes section:
“This review will evaluate all outcomes pertaining to the children’s physical and mental health, as well as cognitive, behavioural and social skills which are presented in studies with the objective of investigating FASD interventions. These outcomes may be measured using standardized/non-standardized and validated/un-validated measures, for example by rating scales (e.g. Child Behaviour checklist). All outcome measures included in the studies will be reported in this review. Special attention will be paid to time of follow up for each outcome, i.e whether outcomes are measured during or immediately after the intervention versus later in life. It is important to note that measures used in this field of study tend to be variable and depend on the intervention targets.

We have provided examples of measures that may be included in studies, however, we do not intended this list to be exclusive and will include trials with any standardized measures, including Eyeberg Child Behaviour Inventory (ECBI), Child Behaviour Checklist (CBCL), Social Skills Rating System (SSRS).
Primary Outcomes:
1. Behavior and social skills: measured by rating skills (e.g. Personal Behavior Checklist scores, Child Behaviour Checklist, Social Skills Rating System);
2. Cognitive abilities: measured by psychometric tests of IQ and memory (e.g. Ballard addition and Subtraction Tests);

3. Educational skills and attainment: measured by grade repetition, special educational supports, and validated scales measuring literacy and mathematical skills (e.g. Phonological Awareness and Early Literacy Test);
Secondary Outcomes:
4. Diagnosis of ADHD: measured by clinical diagnosis and assessment;
5. Psychiatric co-morbidity: measured by rating scales (e.g. Child Depression Inventory, Beck Depression Inventory).
6. Hyperactivity: as measured by rating scales (e.g. Conner's Parent Rating Scale, Conner's Teacher Rating Scale);
7. Impulsivity: as measured by rating scales (e.g. Conner's Parent Rating Scale, Conner's Teacher Rating Scale);
8. Attention: as measured by rating scales (e.g. Conner's Parent Rating Scale, Conner's Teacher Rating Scale)."

In contrast to pharmacological interventions, the adverse outcomes identified for non-pharmacological interventions do not appear to come from separate measures/outcomes than targeted primary/secondary outcomes. But rather, they generally reflect the opposite direction of change that what is targeted. Things like child maltreatment, placement changes, or suicidal ideation/attempts could be considered.
• We agree with the reviewers comments and have such made the following changes on page 9: "new symptoms caused by the behavioral/psychological intervention including, increase in psychiatric symptoms and agitation, as well as possibilities of child maltreatment, and/or suicidal ideation/attempts".

What is the rationale for only reporting the longest follow-up period? It would seem of interest to consider both the immediate post-intervention and longest follow-up period to assess the initial strength of effects and whether effects changed over time (continued to improve, maintain, dissipate). This is particularly relevant for behavioral interventions where behavior change agents (i.e., often parent or teacher) continue to implement (or not) targeted strategies. Also, what will be done in the common situation when there is a delayed-waitlist comparison group who receives the intervention in the interval between immediate post-intervention and subsequent follow-up?
• We agree with the reviewers comments and have edited the sentence on page 14 to read: "For all studies, we will include data from all reported time periods"
• In the case where there may be a delayed waitlist comparison group it would be excepted that studies would ensure that patients do not receive intervention during study follow-up time periods, as this is standard practice in these types of studies, however if studies have not done this we would list it as a limitation in the evaluation of the intervention.

The authors indicate they plan to search conference proceedings from the last 3 years for unpublished results. This reviewer is aware of multiple trials that had strong designs, but have yet to be published dating back to the mid-to late 2000 for various reasons. I'd recommend considering going back further.
The most relevant conferences would be the Research Society on Alcoholism (abstracts published in Alcoholism: Clinical and Experimental Research). And the International Conference on FASD hosted by the University of British Columbia every other year (2017 was 7th conference).
• We have extended going back from the last 5 years for conference proceedings. Going back farther would be increasing our workload for what the team perceives to be of minimal benefit as trials from further than 5 years ago should be published or available to us in the grey literature we are searching. We will be searching both of the above mentioned conference proceedings as mentioned in our search section (page 11).

On page 12, the authors added that a smaller set of criteria will be used to initially assess inclusion based on the abstract and title. This seems appropriate; however, this sentence is vague. What criteria will be used?
• It is common in systematic reviews to only look at the population, intervention and study designs during the title/abstract screening phase – this is because we do not expect all the details of the PICO to be presented in the abstract. We have added that we will be looking at the population, intervention and study design during this phase to provide more detail for our readers and removed the statement "a smaller set of criteria"
• "The reviewers will assess titles/abstracts for studies that meet criteria for: population, intervention and study design". Page 12

The second paragraph on page 12 includes content about the form being piloted that is repetitive from the prior paragraph.
• We removed this repetition.

On page 14, the authors state they will contact study authors for missing data. Do they mean missing data from participants (e.g., due to attrition), unpublished data on other outcomes, or statistics not reported in the paper needed for effect size calculations? They also stated that they will solicit unpublished data from authors in progress or recently completed trials. It seems unlikely researchers will feel comfortable sharing in progress or recently completed data before they have had the chance to present or publish on the data themselves first.
• When conducting a systematic review, review authors traditionally contact authors of included studies for: missing data (for example, SD, data lost due to attrition), or statistics or outcomes that are needed for possible meta-analysis, or results of recently completed, unpublished trials. We have added these details to our missing data section on page 14. The practice of contacting authors helps with decreasing publication bias, and this process is standardized practice for systematic reviews to provide authors with an opportunity to present unpublished data. See Cochrane Review handbook: http://methods.cochrane.org/sites/default/files/public/uploads/mecir_printed_booklet_final.pdf.

In response to reviewer 2's final question, the authors stated they will not be reviewing the quality of the evidence. However, in a prior comment they stated they will be assessing the quality of the evidence when commenting on the validity of measurement scales. The first sentence in the discussion also alludes to using a standardized rating tool to assess methodological quality of studies (assuming this is referring to the bias tool?). This is inconsistent.
• Thank you for picking up on this inconsistency. We will be assessing the risk of bias in individual studies and have edited the first sentence in our discussion to reflect this (page 15). We will also be reviewing the overall quality of the evidence using GRADE and have provided the following information: page 14:
"Grading the Evidence for Each Primary Outcome:
The strength of evidence for the primary outcomes will be graded by using the approach described by the GRADE working group. Two reviewers will evaluate the strength of a body of evidence independently, and discrepancies will be resolved through consensus.

This approach assesses the evidence based on four domains: risk of bias, inconsistency, indirectness, imprecision, publication bias, and other factors (and upgrading). We will classify the strength of evidence as "high", "moderate", "low", or "very low" and make recommendations for future research needs.

The authors have decided to only include studies using an RCT design, which omits some rigorous and well-done studies (e.g., O'Connor et al., 2006 – used alternating assignment within cohorts). Most "real-world" effectiveness in community settings would also be excluded with this criterion, which

some may argue is even more important than RCTs in more controlled laboratory settings. In addition, the authors state they intend to put forth “minimal scientific criteria to guide the next generation of trials in this area to provide more valid and clinically useful information.” This seems to imply that the current body of work is inadequate (despite being small) and that studies using alternate designs are not worthy of consideration. While criteria could be theoretically useful, this reviewer has some concerns that such criteria could hamper future growth of intervention research rather than promote the next generation. Conducting RCTs of interventions for FASD should be encouraged and promoted, but there are limitations to this design in some regards and practical challenging realities (e.g., funding, access to large sample sizes). Progress on empirically validated interventions for people with FASD was almost nonexistent for decades and has only shown gradual growth over the last 10 years. If the authors decide to advocate for minimal scientific criteria, they may want to consider a diversity of designs that promote a logical and systematic approach to intervention development and evaluation. Olson et al., 2016 (ACER) may be useful in considering key research questions for intervention research and potential designs to address these key questions. A diversity of acceptable designs also provides greater opportunities for a broader range of researchers and clinicians to contribute to the scientific enterprise.

• We agree with the reviewer’s comments that our inclusion of only studies using RCT design would omit real world, well-done studies and value the contribution of observational studies in this field. As a first step, we would like to keep the scope of this review to RCT as a first step to assess the evidence form these types of studies and to highlight gaps. Highlighting the gaps may actually further the rational for observational studies to be conducted. We are also undertaking future work that pertains to the review of studies utilizing observational designs and have included this in our discussion.

• We have also included in our limitations the importance of observational studies in this area: “A limitation of this review is the exclusion of studies using observational study designs, as these types of studies are common when assessing treatments for FASD in community settings. Therefore this review may be missing effective FASD interventions. Future work is being conducted by our study team that will expand our program of research to include a systematic review that summarizes studies utilizing observational designs to evaluate FASD interventions.” (page 16).

• We have also taken out our statement to put forth “minimal scientific criteria to guide the next generation of trials” as our intention is indeed to promote the next generation of intervention research in FASD.

Finally, there are several locations where a word or punctuation is missing. A thorough edit is recommended.

• Thank you for your comments, we have edited our paper.

Reviewer 2

SEARCH STRATEGY

The authors have addressed all areas raised in the previous peer-review. The use of forward citation searching for ‘key’ studies is somewhat vague and could introduce bias if the operationalisation of ‘key studies’ is not made explicit in the protocol. I suggest the authors conduct forward citation searching for all studies deemed eligible for review.

• Thank you for your comments. We agree that our current statement is vague and have edited it to state that we will be conducting a forward citation search for all studies that are eligible for review. “We will also perform forward searches of all studies included in this review in Web of Science to identify additional citations that might have been missed in the database search” page 11

INCLUSION CRITERIA/STUDY SELECTION

1. Title/Abstract Screening. The authors have identified that a reduced number of criteria will be used to screen titles/abstracts. Could the authors provide some more explicit detail for what criteria will be used? This may require some piloting. It can be difficult to unequivocally decide on inclusion/exclusion

based on titles and abstracts, so it would be helpful to understand exactly how the authors will conduct this stage of their review in order to assess whether their decided method is sound.
• We agree that our statement “we will use a reduced number of criteria” is vague, therefore we have taken steps to clarify and provide more detail. It is common in systematic reviews to only look at the population, intervention and study designs during the title/abstract screening phase. We have added that we will be looking at the population, intervention and study design during this phase to provide more detail for our readers and removed the statement “a smaller set of criteria”
“The reviewers will assess titles/abstracts for studies that meet criteria for: population, intervention and study design”. Page 12

2. Intervention Setting. If all types of intervention settings are to be included in the review, the authors need to explicitly state this in the protocol.
• We have included the following statement on page 8 “Types of Interventions: All pharmacological or non-pharmacological interventions targeting the improvement of FASD symptoms in children in all types of intervention settings will be included”

3. FASD Diagnosis. Requiring a formal FASD diagnosis could be problematic and unnecessarily exclude some research. Some authors may not explicitly report whether the diagnosis is official or if there even has been a diagnosis (e.g., authors may just state they are using a cohort of children with FASD without any detail on diagnosis). I would suggest including studies that state they are using an FASD sample and studies that report directly on diagnosis. The authors could then code whether or not there is an official diagnosis and see if results differ during the analysis phase.
• Thank you for your helpful suggestions. We agree with you that obtaining data on whether the diagnosis of FASD is “official” is difficult. We have taken your suggestion and have edited the following statement on page 8 to read:
“Types of Participants: Children (<18 years), both males and females, with an author defined diagnosis of FASD (i.e studies which state they are using a cohort of children with FASD) including but not limited to: Fetal Alcohol Syndrome, Partial Fetal Alcohol Syndrome, Alcohol-related Neurodevelopmental Disorder, and Alcohol-related Birth Defects.”

4. Outcomes. It is till not clear whether validated and unvalidated measures are included. The authors’ response to this point is not clear and some of the 'validated' terminology has been removed (e.g., in the abstract), but in other areas it has not (e.g., outcomes section). Moreover, it is still not clear by what the authors mean by 'validated'. How exactly will this be operationalised when making inclusion/exclusion criteria?
Is reliability also important? What type of validity? If this data is not reported in the study, will the study be excluded or will the authors verify psychometric qualities by looking at broader literature for the measure? The authors have also not sufficiently addressed the point on standardised cognitive outcomes. To provide a comprehensive review, I would suggest including both standardised/non-standardised and validated/unvalidated measures. As the authors correctly state, the quality of the measures can be evaluated during the risk of bias assessment. Moreover, the type/quality of outcome measures could be used in a subgroup analysis to examine whether quality of outcome measures impacts the overall impact of interventions.
• Thank you for your helpful suggestions. We will evaluate all outcomes presented in all included studies, we are not excluding studies or outcomes based on the use of validated or unvalidated/ standaridized or non-standardized measures.
• We have stated in the outcome measures section we will include studies that are pertaining to the children’s physical and mental health, as well as cognitive, behavioural and social skills which are presented in included studies. These outcomes may be measured using standardized/non-standardized and validated/un-validated measures, for example by rating scales (e.g. Child Behaviour checklist). All outcome measures included in the studies will be reported in this review. We have

added this to the outcomes section and have removed any specific statements to only validated or standardized measures.

ANALYSIS AND SYNTHESIS

1. Unit of analysis. I am curious as to why the authors would only include the longest follow-up time-point. It would be useful to examine whether effects are seen at post-intervention and then maintain at follow-up time-points. This can be achieved by conducting separate meta- analyses at specific time-points. Perhaps the authors could examine some recent Cochrane reviews that have taken this approach. Examining the impact of interventions over time is an important practice and policy issue. Practitioners and policy-makers need to understand if effects take some time to appear, appear immediately, reduce over time, or improve over time.

• We agree with the reviewers comments and have edited the sentence on page 13 to read: “For all studies, we will include data from all reported time periods”

2. Coding of intervention types for analysis. Could the authors please provide their General approach for how non-pharmacological interventions will be grouped? This could be quite general and informed by previous reviews, but does require some attention so that it can be assessed as part of the overall review protocol (e.g., categorisation by modality or anticipated theoretical underpinning/approach).

• Page 8 “Non-pharmacological interventions will be grouped according to categorization by type of intervention, i.e behavioral intervention, educational intervention, social intervention”.

3. The authors have not addressed the first point on how they will synthesise the evidence if meta-analysis is not appropriate. Moreover, I am curious as to why the authors will not assessing the overall quality of the evidence when this is now considered an important component of systematic reviews. Could the authors please provide a justification for this or include their plans for conducting an appraisal of the body of evidence (latter more acceptable approach).

• If meta-analysis is not appropriate, we will conduct a quailative/narrative synthesis. As for narrative synthesis, there is limited guidance or standardization outside the conduct of a formal qualitative synthesis. It has been reported that while narrative synthesis are used in as much as half or more of systematic reviews in certain areas, their reporting is often inadequate (Campbell, Lancet, 2016). We will use guidance provided by several organizations in the conduct/ reporting of a narrative summary (if needed). Examples of the guidance documents including Cochrane (Ryan R; Cochrane Consumers and Communication Review Group.

‘Cochrane Consumers and Communication Review Group: data synthesis and analysis’. http://cccrg.cochrane.org, June 2013), Lancaster University (Popay et al, Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product from the ESRC Methods Programme, 2006) and the NIHR Complex Reviews Support Unit (http://www.nihrcrsu.org/guidance/narrative_synthesis/).

• We are going to evaluate the strength of the overall evidence using GRADE and have include this in our protocol on page 14:

“Grading the Evidence for Each Primary Outcome:

The strength of evidence for the primary outcomes will be graded by using the approach described by the GRADE working group. Two reviewers will evaluate the strength of a body of evidence independently, and discrepancies will be resolved through consensus. This approach assesses the evidence based on four domains: risk of bias, inconsistency, indirectness, imprecision, publication bias, and other factors (and upgrading). We will classify the strength of evidence as “high”, “moderate”, “low”, or “very low” and make recommendations for future research needs.

VERSION 3 – REVIEW

REVIEWER	Elizabeth Eggins University of Queensland Australia
REVIEW RETURNED	13-Oct-2017

GENERAL COMMENTS	The authors have provided a satisfactory response to the previous peer-review comments and an improved systematic review protocol. Again, I appreciated their clear outline of how they have addressed each comment from the initial peer-review. There are still some minor outstanding issues that require addressing before the protocol is ready for publication. The points of revision are provided below. INCLUSION CRITERIA/STUDY SELECTION 1. Title/Abstract Screening. While it is often common in more medical focused systematic reviews to include study design in the title and abstract screening stage, this can be problematic when reviews include research from other disciplines. This is because abstracts can be unstructured and not include the level of detail required to unequivocally exclude a title/abstract. If the authors will include any records where it is not unequivocally clear if the criterion has been satisfied, then the approach is sound. However, this needs to be specified in the protocol. 2. Primary and Secondary Outcomes. Can the authors please specify if a study needs to have reported the primary outcome to be included and only then will secondary outcomes be coded. The authors need to be cautious of this approach because it means that any synthesis of secondary outcomes is not a comprehensive synthesis of the literature because it will only provide a synthesis of studies with a primary outcome AND a secondary outcome. It will not include studies that have only a secondary outcome. This could have implications for the conclusions and practice recommendations that are derived from the review. ANALYSIS AND SYNTHESIS 1. Unit of analysis. The issue regarding follow-up time-periods has not been sufficiently addressed. For example the following statements are not sufficient for understanding how the authors will handle studies with multiple time-points: • "Special attention will be paid to time of follow up for each outcome, i.e whether outcomes are measured during or immediately after the intervention versus later in life" – what do the authors mean by "special attention"? • "For all studies, we will include data from all reported time periods" – good decision, but HOW will the outcomes across time-points be synthesised? (see suggestions in previous review for addressing this)

VERSION 3 – AUTHOR RESPONSE

Reviewer 3:

INCLUSION CRITERIA/STUDY SELECTION

1. Title/Abstract Screening. While it is often common in more medical focused systematic reviews to include study design in the title and abstract screening stage, this can be problematic when reviews include research from other disciplines. This is because abstracts can be unstructured and not include

the level of detail required to unequivocally exclude a title/abstract. If the authors will include any records where it is not unequivocally clear if the criterion has been satisfied, then the approach is sound. However, this needs to be specified in the protocol.
- Thank you for your comment. At this stage, we will not exclude citations on the basis of them not being described as being randomized. Having said, for a citation to be included, the authors have to describe a comparative study. This is now specifically stated in our manuscript on page 11:
"At this stage, authors will not exclude citations on the basis of them not being cited as randomized, however for a citation to be included, the authors have to describe a comparative study".

2. Primary and Secondary Outcomes. Can the authors please specify if a study needs to have reported the primary outcome to be included and only then will secondary outcomes be coded. The authors need to be cautious of this approach because it means that any synthesis of secondary outcomes is not a comprehensive synthesis of the literature because it will only provide a synthesis of studies with a primary outcome AND a secondary outcome. It will not include studies that have only a secondary outcome. This could have implications for the conclusions and practice recommendations that are derived from the review.
- Thank you for comment. We agree and were planning on including studies that included any of our primary and secondary outcomes. We will not exclude studies on the basis of outcomes. If any of the primary or secondary outcomes are reported then the study is eligible for inclusion (if there is no other reason for exclusion). We have stated this explicitly in our review on page 8:
"Furthermore, studies will not be excluded on the basis of outcomes, if any of the primary or secondary outcomes are reported then the study is eligible for inclusion."

ANALYSIS AND SYNTHESIS
1. Unit of analysis. The issue regarding follow-up time-periods has not been sufficiently addressed. For example the following statements are not sufficient for understanding how the authors will handle studies with multiple time-points:
"Special attention will be paid to time of follow up for each outcome, i.e whether outcomes are measured during or immediately after the intervention versus later in life" – what do the authors mean by "special attention"?
"For all studies, we will include data from all reported time periods" – good decision, but HOW will the outcomes across time-points be synthesised? (see suggestions in previous review for addressing this)
• Thank you for your comments. We have removed the statement "special attention…."as we agree this is vague and replaced it with:
"Follow-up data will be collected from all reported time periods" – page 8
• Furthemore, we have provided instructions on how we are going to synthesize outcoms from across time points on page 14:
"For all studies, we will include data from all reported time periods; separate meta-analyses will be conducted for outcomes measured immediately after intervention, 6 months after, and > 12 months after the intervention".