Dear Dr Paul Whaley,

Thank you for providing a detailed Evaluation Report for TEBT-2023-0010 submission **'Evidence on the effects of Flame Retardant Substances at Ecologically Relevant Endpoints: A Systematic Map Protocol**'. We greatly appreciate the time and attention taken by the reviewers to provide detailed feedback during round 2 of peer review, and thank you for the opportunity to revise the protocol. We have considered each comment in detail and have made a series of changes to the protocol. A revised version of the protocol has been uploaded to Zenodo, along with any revised supplementary material. The revised protocol has also been re-submitted via the EBT submission portal. As requested, any amendments and/or revisions to the text have been made in coloured text. Below we document the actions taken in response to each comment. For ease, the reviewers' comments are highlighted in green with our response in black below.

Reviewers comments

General comments

1. Overall the revision is a significant improvement to the protocol and the authors are applauded for the thought and care that has gone into crafting a well planned protocol. It is still not clear to me, however, if the authors have intentions to make more than a filterable excel-based database or if they will commit to also making interactive figures.

Thank you for your kind words on revisions made to the previous protocol. As per the journal policy the primary product of this evidence map will be the interrogable excel-based database published alongside the narrative summary. We also intend to produce interactive visualisations and figures however this will be strictly a secondary output and will balance capacity with complexity of the final database. More detail is provided on this below with revisions and/or further clarity provided in the manuscript.

2. Thank you for your thoughtful responses to the reviewers' comments. I think this a much improved protocol given the changes and clarifications you've made.

Remaining issues that could materially impact the scientific quality of the planned research

3. In general, recommend careful consideration and use of the terms "database" and "map". The authors responded to reviewers that they would not be generating a map in Tableau, which is understandable, but it is unclear what type of "map" the authors might be aiming to create, as it is indicated in Section 2.7 that a "coded, interrogable database (i.e., map)" will be available, but it seems that the authors are referring only to a filterable/sortable excel database (which is provided as an example in Supplemental Material 2.6). While a database is beneficial, it does not provide high level summary information or visualizations that automatically update based on end-user

sorting/selecting of information. The benefit of these tools is for end users to select/filter data and have other outputs (i.e. histograms or charts) update automatically. A filterable excel sheet is nice, but it does not provide this functionality.

We agree with the reviewers that the terms “database” and “map” are not clearly defined in the manuscript and require careful consideration. As per the editors suggestion we have amended the manuscript to refer to the primary output of the exercise to be an “evidence map”. We also agree that interactive visualisation tools will greatly benefit the dissemination of the data and improve end user understanding and experience. Whilst we have chosen not to map the data on Tableau, we intend to visualise the data using an interactive interface such as Microsoft PowerBI or using an R Shiny app. We have provided further clarity in the manuscript on the intended outputs which include a narrative summary and evidence map.

4. Perhaps authors could check out this as an alternative data visualization option (here) with data available (here). Earlier reviewer comments were not to say that the authors had to necessarily use Tableau, but that authors should begin considering how to use various tools to create the type of interactive or interrogable outputs that they are envisioning. Other options include PowerBI, for example: If the authors are not planning to provide more than a filterable excel spreadsheet, then I suggest that the term and phrase “map” not be used to describe the final product, and rather refer to the outcome as a database.

We thank the reviewers for the clarification on earlier comments. We intend to produce an interactive and/or interrogable output for visualising the data as explained in response to point 3 and thank the reviewers for the helpful suggestions. We have provided further clarity in the manuscript on the intended outputs which include a narrative summary and evidence map.

Comments on approach to quality control of data abstraction

5. The data extraction table in 2.5 is very comprehensive and may likely take a lot of time to complete. I see that column BM, BN, and BO are reserved for QC feedback. If the QC indicates changes will be made, will they be instructed to make the changes in the row or will a new row be created? What is the process here?

The quality control (QC) measures described in the data extraction table have been designed to ensure the second reviewer can clearly indicate if they 1. agree or disagree with the inclusion and/or exclusion of an article (column BM), and 2. agree with the reliability of extracted data (column BO). All reasons for disagreement in the inclusion/exclusion and/or reliability of extracted data will be noted in the table (columns BN and BP respectively). For clarity additional columns have been added to separate agreement in inclusion/exclusion and reliability of extracted data.

To assess for consistency (i.e., agreement between reviewers in inclusion/exclusion), a kappa analysis will be performed at the level of full text using a random subsample of 10% of all articles. This is in line with good practice as set out by the Collaboration of Environmental Evidence (CEE). The score

of the kappa analysis will be calculated and discussed in the narrative summary. It is difficult to assess the reliability of extracted data using a formal kappa analysis, however the second reviewer will assess 25% of all included articles to assess human error and/or misinterpretation of the data. Any disagreement in data extraction will be noted in column BP and discussed in the narrative summary.

The second reviewer will also assess reasons for exclusion of all (100%) articles excluded at the level of full text with the reasons for exclusion noted in column O. If the second reviewer deems an article has incorrectly been excluded, both reviewers will discuss and re-assess the article for inclusion. If the reviewers agree the article should be included, data extraction will be undertaken by the primary reviewer using the necessary number of rows in the data extraction database.

Section 2.6 of the protocol has been amended to provide greater clarity on the QA/QC measures taken throughout each stage of review.

6. It is not clear to me if there are best practice guidelines for the amount of QC/QA that needs to be performed on data coding and extraction for a systematic evidence map, but note that the authors have indicated that only 10% of included articles will be checked for QC at the data extraction level.

This protocol has been developed in line with guidelines set out by the Collaboration of Environmental Evidence (CEE). This includes best practice guidelines for QC/QA as explained above.

7. It is not clear to me if there is any value in resolving discrepancies in only the 10% of records for which data extraction is checked, as the other 90% of the errors (assuming even distribution of errors across the dataset) will still be present. It would seem to me to be more efficient to record the error rate in the results section and discuss it in the limitations section of the final manuscript.

We thank the reviewers for this useful comment. As per our response to point 5, we do not intend to resolve any discrepancies in data extraction between reviewers and agree that it would be more efficient to calculate and record the error rate (in data extraction) and kappa score (for consistency) and discuss in the limitations section of the narrative summary.

Other remaining issues with the planned methods

8. Comment 20 in the evaluation report was based on concern about null entries in the "chemical name" column, which could result in several different types of exposure appearing the same to the database user. Something usefully descriptive should entered in place of N/A, even if it is as simple as "other" - N/A has no obvious meaning or semantic content, which seems not right, and null entries in databases are best avoided.

We thank the reviewer for clarification of comment 20 in the initial evaluation report. We have amended both the extraction and coded database to avoid null entries in the chemical name column as well as replacing any mention of NA with more semantic content (i.e, not relevant, no statement, not a mixture etc).

9. The authors should make one final check that there are not different values with the same meaning in the database. For example, in the sample data there is a code "unknown" and a code "not stated", both of which presumably mean the same thing? This could interfere with counts of the code and/or interpretation of those counts.

Thank you for this useful comment. The authors have ensured that values are consistent across both the data extraction database and final coded database. Any values referring to 'unknown' have been replaced with 'not stated' for consistency.

10. Line 353: "Behaviour" is VERY broad category- does this include ANY behavioural analysis - aggression, mate choice, anxiety, hyperactivity, social behaviour, etc? Or will only specific behaviours be included?

We do not intend to place a limit on the inclusion of behavioural effect therefore will include any article that studies the effect of a substance on the behaviour of an organism so long as all criteria for inclusion are met. Dependent on the data (i.e., behaviours) that is extracted from the included articles, we intend to sub-categorise behaviour into groups (i.e., movement, mating, foraging etc) with these groups, and the behaviour therein outlined in the code book.

11. Lines 297-298: Please list out the databases available to you on the WOS platform either in the text or the supplemental material (apologies if it's there and I missed it) so that the searches could be reproduced.

The full list of databases included in the Web of Science University of Sheffield subscription has been listed in supplemental material 2.3. This has been noted in the text of the manuscript.

Other questions and comments about the protocol

12. If there are word limits for the abstract it seems that it could be streamlined a fair bit.

We have streamlined the abstract to meet the guidelines set by the journal for abstract length.

13. Line 88: suggest it should be "pathways of exposure" instead of "pathways to exposure"

We agree and have amended this sentence in the manuscript.

14. Section 2.1: "stakeholder" can have negative connotations, and also doesn't feel like the correct term to identify this group. Perhaps "technical advisors" is an appropriate alternative?

We thank the reviewers for this useful comment. We have amended section 2.1 of the manuscript to remove any mention of “stakeholder” and replace this with “technical expert”.

15. Line 259: Is this earlier version of the “full protocol” available online anywhere and/or can it be/should it be added to the OSF entry so that it can be seen what the stakeholders were commenting on?

The earlier version of the protocol that was shared with technical experts has been uploaded to the OSF site and reference to the supplemental material made in the text of the protocol.

16. Lines 282-285 – this is an incomplete sentence. It is also unclear from the writing whether the QSUR has already been run/conducted and is citable or if this is a step that the authors will also be conducting.

Lines 282-285 have been amended in the manuscript to read as a more complete sentence. We have noted in the manuscript that the QSUR analysis has already been performed by Bevington et al., 2022 prior to inclusion in the Inventory and have included the appropriate citation.

17. Lines 302 and 304 both discuss information sources that will not be included. Suggest keeping this together by stating at line 304 “Grey literature, including academic theses and dissertation databases will not be searched."

We agree and have amended section 2.2 of the manuscript as suggested.

18. Line 307: unclear what "these" are - the number of studies retrieved or the retrieved studies themselves

This sentence has been amended to more clearly note that the spreadsheet and PRISMA flow diagram - previously referred to as ‘these’ - will be included alongside the final manuscript.

19. Line 362: The parenthetical "controlled" isn't in Table 3. It is unclear to me the impact of the parenthetical in terms of studies that are ultimately included or not.

We agree with the reviewer that the inclusion of “controlled” in this sentence is unnecessary and have removed it. We detail the inclusion/exclusion criteria for the comparator (i.e., control group) later in the manuscript.

20. Suggest deleting the final sentence in the paragraph ending in Lines 404 and 405 as it is redundant with the start of the next paragraph.

We agree and this sentence has been removed.

21. Lines 404-407: Could this also be stated: "All articles will be screened by one reviewer at the title and abstract stage with Rayyan serving as the second reviewer."? Will the 1 reviewer be aware of the machine learning decision on the paper when they review?

We do not agree that this statement should be added to the protocol as we do not agree that Rayyan can or indeed will serve as a second reviewer. Using the keyword and PECO highlighting tools it is possible for Rayyan to 'learn' and thus compute a relevance rating score of all remaining articles for inclusion. Rayyan does not provide a yes/no decision on whether to include/exclude an article, it instead provides a numerical rating score of whether it considers an article relevant based on the PECO criteria. It will be possible for the primary reviewer to order the articles for review in order of relevance. As noted in the protocol all articles will be assessed for inclusion by the primary reviewer with Rayyan's machine learning functions used primarily to reduce time when screening. We do not consider Rayyan to act as a second reviewer, and thus do not agree that this statement can be made.

22. Note in supplemental material 2.6 the citation format is not consistent for the three piloted references.

Thank you for this useful comment - we have amended the citations to all follow the APA format.