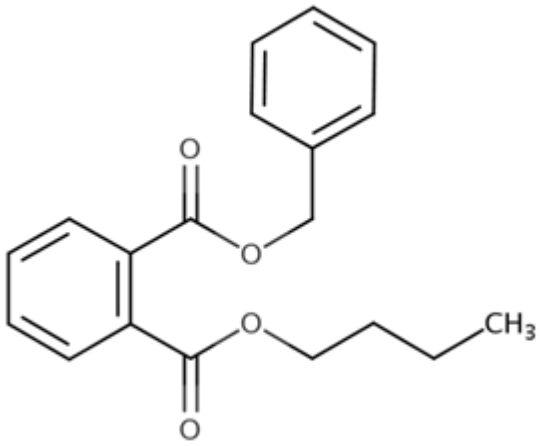
United States
Environmental Protection Agency

December 2025
Office of Chemical Safety and
Pollution Prevention

Systematic Review Protocol for Butyl Benzyl Phthalate (BBP)

Systematic Review Support Document for the Risk Evaluation

CASRN 85-68-7



December 2025

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The U.S. EPA's Office of Pollution Prevention and Toxics (OPPT) applies systematic review principles in the development of risk evaluations under the amended Toxic Substances Control Act (TSCA). TSCA section 26(h) requires EPA to use scientific information, technical procedures, measures, methods, protocols, methodologies, and models consistent with the best available science and base decisions under Section 6 on the weight of the scientific evidence. Within the TSCA risk evaluation context, the weight of the scientific evidence is defined as "a systematic review method, applied in a manner suited to the nature of the evidence or decision, that uses a pre-established protocol to comprehensively, objectively, transparently, and consistently identify and evaluate each stream of evidence, including strengths, limitations, and relevance of each study and to integrate evidence as necessary and appropriate based upon strengths, limitations, and relevance" (40 CFR 702.33).

To meet the TSCA section 26(h) science standards, EPA used the TSCA systematic review process described in the *Draft Systematic Review Protocol Supporting TSCA Risk Evaluations for Chemical Substances* (U.S. EPA, 2021) (hereinafter referred to as "2021 Draft Systematic Review Protocol"). Section 3 of the 2021 Draft Systematic Review Protocol depicts the steps in which information is identified and whether it undergoes the formal systematic review process (U.S. EPA, 2021). Information attained via the systematic review process is integrated with information attained from sources of information that do not undergo systematic review (*e.g.*, EPA-generated model outputs) to support a weight of the scientific evidence analysis.

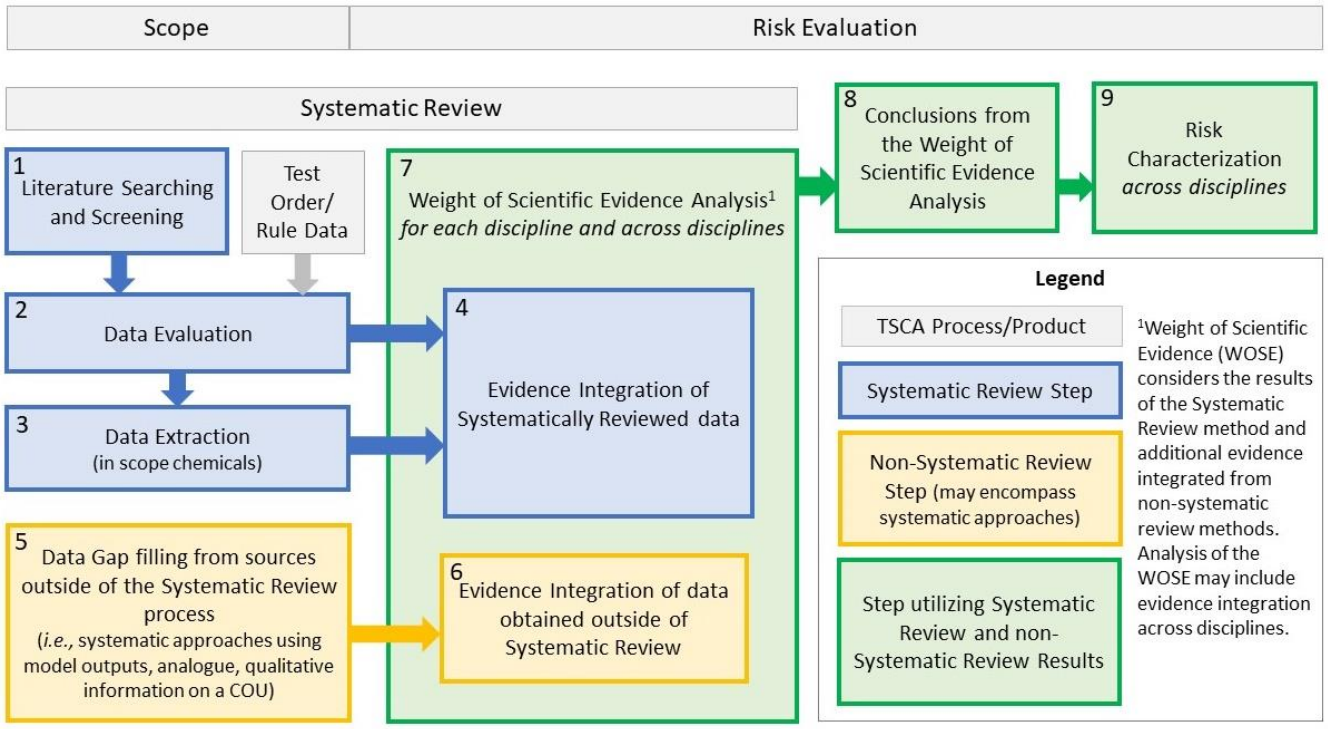


Figure 1-1. Overview of the TSCA Risk Evaluation Process with Identified Systematic Review Steps

The process complements the risk evaluation process in that it is used to develop the exposure and hazard assessments based on reasonably available information. EPA defines "reasonably available information" to mean information that EPA possesses or can reasonably obtain and synthesize for use in risk evaluations, considering the deadlines for completing the evaluation (40 CFR 702.33).

2 CLARIFICATIONS AND UPDATES TO THE 2021 DRAFT SYSTEMATIC REVIEW PROTOCOL

In 2021, EPA released the *Draft Systematic Review Protocol Supporting TSCA Risk Evaluations for Chemical Substances* (U.S. EPA, 2021), a framework of systematic review approaches under TSCA, to address comments received on a precursor systematic review approaches framework, the *Application of Systematic Review in TSCA Risk Evaluations* (U.S. EPA, 2018). In April 2022, the SACC provided comments on the 2021 Draft Systematic Review Protocol and additional comments on OPPT's systematic review approaches were garnered during the public comment period. In lieu of an update to the 2021 Draft Systematic Review Protocol, this systematic review protocol for the *Risk Evaluation for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025q) (hereinafter referred to as "Risk Evaluation for BBP") describes some clarifications and different approaches that were implemented than those described in the 2021 Draft Systematic Review Protocol in response to (1) SACC comments, (2) public comments, or (3) to reflect chemical-specific risk evaluation needs.

2.1 Clarifications

The chemical-specific systematic review protocol is used to transparently document any updates or clarifications made to the systematic review process used for considering information identified for a given TSCA risk evaluation, as compared to those published in the Draft Systematic Review Protocol Supporting TSCA Risk Evaluations for Chemical Substances (U.S. EPA, 2021). Throughout the 2021 Draft Systematic Review Protocol, there were some terms used that were not explicitly defined, resulting in their different uses within the document (U.S. EPA, 2021). Table 2-1 lists the terms that were updated to resolve some of the confusion expressed by the public and SACC comments regarding the implementation of the respective systematic review-related step. One main clarification is that *all references that undergo systematic review are considered for use in the risk evaluation*, even those that do not meet the various discipline and sub-discipline screening criteria or those that are categorized as supplemental information at title and abstract (TIAB) or full-text screening.

Section 4.2.5 of the 2021 Draft Systematic Review Protocol describes how data sources (*e.g.*, individual references, databases) may be tagged and linked in when the same information is present in multiple publications (U.S. EPA, 2021). References will generally undergo data quality evaluation and extraction if there are data that pass screening criteria; however, to prevent the same data from being represented multiple times and conflating the amount of available information there is on a subject area, if two or more references contain the same results tables, EPA selects the reference(s) that most thoroughly describes the extractable results (indicated as the parent reference in DistillerSR). If two references portray the same information from the same dataset, only one is counted in the overall dataset (*i.e.*, deduplication). If two references contain information about the same dataset, but one of those references only provides additional contextual information or summary statistics (*e.g.*, mean), both data sources are linked but the extractable information from both may be combined in DistillerSR. This enables the capture of key information while avoiding double-counting the data of interest. The linked reference containing most of the data, which are evaluated and extracted, is identified in DistillerSR as the parent reference; the "complementary child reference" in DistillerSR does not undergo independent data evaluation and extraction but is evaluated and extracted in combination with the parent reference. Linking the references in DistillerSR allows the reference with more limited information or only contextual information to be tracked and utilized to evaluate the extracted data in the other related studies. The child reference may undergo data quality evaluation and extraction if there are additional unique and original data that pass screening criteria.

Section 4.5 of the 2021 Draft Systematic Review Protocol describes how data may be obtained using TSCA authorities and test orders. One update to that section is that in addition to requiring data reporting under TSCA sections 4 (test order), 8(a) (Chemical Data Reporting) and 8(d) (Health and Safety Data Reporting), *EPA may also require data reporting under TSCA section 8(c)* (Call-in of Adverse Reactions Records). Appendix 5.3 also describes how information may be submitted to EPA under other TSCA authorities (*e.g.*, TSCA sections 4, 5, 6, 8(d) and 8 (e), as well as FYI submissions).

Section 5 of the 2021 Draft Systematic Review Protocol describes how EPA conducts data quality evaluation of data/information sources considered for a respective chemical risk evaluation, with Section 5.2 specifically explaining the terminology used to describe both metric and overall data/information source quality determinations (U.S. EPA, 2021). To respond to both SACC and public comments regarding the inappropriate use of quantitative methodologies to calculate both "metric rankings" and "overall study rankings," *EPA decided to not implement quantitative methodologies to attain either metric and overall data/information source quality determinations* and therefore updated the terminology used for both metric ("metric ranking") and overall data/information source ("overall study ranking") quality determinations (Table 2-1). Subsequently terminology for both individual metric and overall information source quality determinations has been updated to "metric rating" and "overall quality determination," respectively. The word "level" was also often used synonymously and inconsistently with the word "ranking" in the 2021 Draft Systematic Review Protocol; that inconsistency has been rectified, resulting in the word "level" no longer being used to indicate either metric or overall data/information source quality determinations (U.S. EPA, 2021).

Sections 4.3.2.1.3 and 6 of the 2021 Draft Systematic Review Protocol describe when EPA may reach out to authors of data/information sources to obtain raw data or missing elements that are important to support the data evaluation and data integration steps (U.S. EPA, 2021). In such cases, the request(s) for additional data/information, number of contact attempts, and responses from the authors are documented. EPA's outreach is considered unsuccessful if those contacted do not respond to email or phone requests within one month of initial attempt(s) of contact. One important clarification to this guidance is that *EPA may reach out to authors anytime during the systematic review process for a given data/information source or reference, and that contacting authors does not explicitly happen during the data quality evaluation or extraction step.*

Table 2-1. Terminology Clarifications between the 2021 Draft Systematic Review Protocol and the Risk Evaluation for BBP

2021 Draft Systematic Review Protocol Term	BBP Systematic Review Protocol Term Update	Clarification
"Title and abstract" or "title/abstract"	"Title and abstract"	To increase consistency, the term "title and abstract" will be used to refer to information specific to "title and abstract" screening.
Variations of how "include," "on topic" or "PECO[a]/PESO[b]/RESO[c] relevant" implied a reference was considered for use in the risk evaluation, whereas	Meets/does not meet PECO[a]/PESO[b]/RESO[c] screening criteria	The term "include" or "exclude" falsely suggests that a reference was or was not, respectively, considered in the risk evaluation. There was also confusion regarding whether "on topic" and "PECO[a]/PESO[b]/RESO[c] relevant" were synonymous and suggested those references were explicitly considered for use in the risk evaluation (and by

2021 Draft Systematic Review Protocol Term	BBP Systematic Review Protocol Term Update	Clarification
"exclude," "off topic" or "not PECO[a]/PESO[b]/RESO[c] relevant" implied a reference was *not* considered for use in the risk evaluation.		default, "off topic" and "not PECO[a]/PESO[b]/RESO[c] relevant" references were not). References that meet the screening criteria proceed to the next systematic review step; however, all references that undergo systematic review at any time are considered in the risk evaluation. Information that is categorized as supplemental or does not meet screening criteria are generally less relevant for quantitative use in the risk evaluation but may be considered if there is a data need identified. For instance, mechanistic studies are generally categorized as supplemental information at either title and abstract or full-text screening steps but may undergo the remaining systematic review steps if there is a relevant data need for the risk evaluation (*e.g.*, dose response, mode of action).
Database source not unique to a chemical	Database	Updated term and definition of "Database": Data obtained from databases that collate information for the chemical of interest using methods that are reasonable and consistent with sound scientific theory and/or accepted approaches and are from sources generally using sound methods and/or approaches (*e.g.*, state or federal governments, academia). Example databases include STORET (STOrage and RETrieval) and the Massachusetts Energy and Environmental Affairs Data Portal. The term in the 2021 Draft Systematic Review Protocol (Table_Apx N-1) incorrectly suggested that databases that contain information on a singular chemical are not considered (U.S. EPA, 2021). Furthermore, the wording "large" was removed to prevent confusion and the incorrect suggestion that there is a data size requirement for databases that contain information that may be considered for systematic review.
Metric Ranking or Level	Metric Rating	As explained above, EPA is not implementing quantitative methodologies to indicate metric quality determinations, therefore the term "ranking" is inappropriate. The term "level" was inconsistently used to indicate metric quality determinations previously; therefore, EPA is removing the use of this term to reduce confusion when referring to metric quality determinations. The term "Rating" is more appropriate to indicate the use of professional judgement to determine a quality level for individual metrics.

2021 Draft Systematic Review Protocol Term	BBP Systematic Review Protocol Term Update	Clarification
Overall Study Ranking or Level	Overall Quality Determination (OQD)	As explained above, EPA is not implementing quantitative methodologies to indicate overall data/information source quality determinations, therefore the term "ranking" is inappropriate. The term "level" was inconsistently used to indicate overall data/information source quality determinations previously; therefore, EPA is removing the use of this term to reduce confusion when referring to overall data/information source quality determinations. The term "Rating" is more appropriate to indicate the use of professional judgement to determine a quality level for the overall data/information source quality determination.
Sub-discipline	No change in term	Sub-discipline explicitly indicates the two categories of receptor-based studies relevant to evaluate human health hazard (discipline): epidemiological (human receptor) or human health animal model toxicological studies (non-human animal receptor). Although environmental hazard is a discipline, Appendix T incorrectly suggested that environmental hazard is a sub-discipline in the 2021 Draft Systematic Review Protocol.
Evidence Stream	No change in term	Evidence streams were updated for both environmental and human health hazard disciplines to more appropriately categorize the hazardous endpoints that were considered. Please see additional descriptions of the evidence stream updates in Section 6.5 below.

[a] "PECO" stands for Population, Exposure, Comparator or Scenario, and Outcomes.
[b] "PESO" stands for Pathways or Processes, Exposure, Setting or Scenario, and Outcomes.
[c] "RESO" stands for Receptors, Exposure, Setting or Scenario, and Outcomes.

3 DATA SEARCH

As described in Section 4 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), EPA conducts a comprehensive search for reasonably available information to support the TSCA risk evaluations. Chemical-specific literature searches are conducted as described in Section 4.2.1 of the 2021 Draft Systematic Review Protocol for all disciplines (*i.e.*, physical and chemical properties, environmental fate and transport properties, engineering, exposure, environmental hazard, and human health hazard) (U.S. EPA, 2021). Additional details on the chemical verification process, and the methodology used to search for chemical specific peer-reviewed and gray literature is available in Sections 4.2 and 4.3 of the 2021 Draft Systematic Review Protocol, respectively (U.S. EPA, 2021). The search for peer-reviewed and gray literature relevant references was completed in September and May 2019, respectively. Appendix Section C.1.17 contains the specific search strings used to identify peer-reviewed literature on BBP (U.S. EPA, 2021). All reasonably available information submitted to EPA under TSCA authorities was considered.

3.1 Multi-Disciplinary Updates and Clarifications to the Data Search

For the Risk Evaluation for BBP (U.S. EPA, 2025q), the literature search in 2019 was conducted as described in Section 4 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), where the peer-reviewed and gray literature updated search followed the approach outlined in Sections 4.2 and 4.3 of the 2021 Draft Systematic Review Protocol, respectively (U.S. EPA, 2021). Occasionally additional data sources relevant for the risk evaluation may be identified after the initial search for peer-reviewed and gray literature; these data sources will then undergo systematic review for the relevant discipline(s). Additionally, each discipline utilizes different strategies (*e.g.*, search strings) to attain their discipline-specific pools of data sources that undergo systematic review.

SWIFT-Review Validation

EPA received comments regarding the lack of detail on the use and validation of SWIFT-Review to determine discipline-specific peer-reviewed reference set considered for use in TSCA risk evaluations. In response to those comments, EPA conducted validation exercises to clarify the search process and build consistency among all the disciplines. The 2021 Draft Systematic Review Protocol contains validation results for the use of SWIFT-Review to determine which peer-reviewed references may be relevant for the characterization of occupational exposure and environmental releases and general population, consumer, and environmental exposure for the respective chemical risk evaluations. However, to expand upon the information provided in the 2021 Draft Systematic Review Protocol, EPA validated references relevant for determining chemical-specific peer-reviewed reference set for the characterization of physical and chemical properties, environmental fate and transport properties, and environmental and human health hazard. EPA manually screened the references that were found in the overall peer-reviewed search results that did not undergo TIAB screening (*i.e.*, references that were not identified using a discipline-specific search string). If a reference that did not undergo further review after TIAB screening was found to meet the screening criteria for a respective discipline (*e.g.*, data needs on physical chemical properties, environmental fate and transport properties, and environmental and human health hazard) and identified for the chemical of interest, it was flagged as a false negative. This analysis validated and verified the use of the search terms in SWIFT-Review, as it showed that less than five percent of references were false negatives across all three disciplines. This method was repeated for several of the TSCA High Priority Substances to build confidence in our discipline-specific search strings.

Supplemental Filtering of 2019 Literature Search for Dermal Absorption
Dermal absorption studies are needed to accurately assess dermal exposure associated with specific conditions of use. Typically, dermal absorption studies are identified as supplemental studies within the human health hazard discipline using the hazard PECOs presented in Appendix H of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). However, dermal absorption data may not meet the screening criteria for other disciplines; these criteria are also presented in Appendix H of U.S. EPA (2021).

To identify any additional studies not found during hazard screening that might be potentially relevant for characterizing dermal absorption and exposure, EPA developed a key word list (identified as a search string in Section 3.7.1 below) and used SWIFT-Review to search/filter the data sources that were previously identified in the BBP chemical search conducted in 2019. EPA followed processes described in the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021): Section 4.2.2 outlines when EPA uses supplemental searching and filtering; and Section 4.2.4 presents the process of using SWIFT-Review to filter data sources identified in the initial chemical search.

Additional Gray Literature Sources
Physical and Chemical Properties: In addition to the gray literature sources listed in Appendix E of the 2021 Draft Systematic Review Protocol, an additional database was added to the list of gray literature sources for physical and chemical properties. The National Institutes for Standards and Technology (NIST) Chemistry Webbook was searched in September 2021 to capture spectroscopic data, specifically ultra-violet and visible absorption (UV-Vis) data, if recorded. This source may also provide thermodynamic data that informs chemical stability and behavior under various conditions.

General Population, Consumer, and Environmental Exposure: In addition to the gray literature sources listed in Appendix E of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), additional sources were added in 2023 and later to capture database outputs from several governmental sources. All datasets were accessed directly and uploaded into HERO. EPA downloaded data from the Centers for Disease Control (CDC) and Prevention's National Health and Nutrition Examination Survey (NHANES). The other datasets included a technical report on human biomonitoring of environmental chemicals in Canada which was conducted by the Government of Canada along with an earlier report by Health Canada.

To obtain information on BBP exposures to the U.S. population, EPA added data from the Centers for Disease Control and Prevention's National Health and Nutrition Examination Survey (NHANES) to its literature set. Although NHANES did not contain relevant information on BBP, EPA did identify potentially relevant information on its primary metabolite, Mono-benzyl phthalate (MBzP). After entering the human body, BBP is metabolized into MBzP in urine. NHANES data on MBzP was also evaluated as part of the systematic review process for data on general population, consumer, and environmental exposure. At the time of download, the three tables available from CDC included "Analysis of Whole Blood, Serum, and Urine Samples, NHANES 1999-2018," "Analysis of Pooled Serum Samples for Select Chemicals, NHANES 2005-2016," and "Analysis of Chemicals Found in Cigarette Smoke in a Special Sample of U.S. Adults, NHANES 2011-2016." Of these, the only dataset containing MBzP data was "Analysis of Whole Blood, Serum, and Urine Samples, NHANES 1999-2018." and the relevant NHANES data were also uploaded into HERO.

New Literature Update
To update the literature pool to capture and consider information published since the original literature search was conducted in 2019 for BBP, EPA identified additional references submitted to the Agency by

the SACC during the peer review of the *Draft Risk Evaluation for Butyl Benzyl Phthalate (BBP)* as well as additional references submitted during the public comment period of the risk evaluation between August 6, 2025 and October 6, 2025. EPA reviewed the list of submitted data sources and identified those that were within scope and had not already been identified and proceeded to screening these data sources as described in Section 4.

3.2 Physical and Chemical Properties

The search for peer-reviewed and gray literature are as described in Sections 4.2 and 4.3, respectively, in the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). SWIFT-Review was used to identify peer-reviewed references that are predicted to be the most relevant for evaluating physical and chemical properties for BBP. Specifically, the search string used to identify data sources that potentially contain physical and chemical property information on BBP in SWIFT-Review was developed by EPA's ORD in collaboration with Sciome and is presented in Appendix G, Section G-1, Table_Apx G-1 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). As mentioned above in Section 3.1, the search string used to identify potentially relevant peer-reviewed data references for evaluation of the physical and chemical properties of BBP was validated. When the search string terms are identified in the title, abstract or as a keyword of a given reference in SWIFT-Review, those references proceed with title and abstract screening.

3.3 Environmental Fate and Transport Properties

The search for peer-reviewed and gray literature are as described in Sections 4.2 and 4.3, respectively, in the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Specifically, SWIFT-Review was used to identify peer-reviewed references that are predicted to be the most relevant for evaluating environmental fate and transport properties for BBP The search string used for environmental fate and transport literature in SWIFT-Review was developed by EPA's ORD in collaboration with Sciome and is presented in Appendix G, Section G.2, Table_Apx G2 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). As mentioned above in Section 3.1, the search string used to identify potentially relevant peer-reviewed data references for evaluation of the environmental fate and transport properties of BBP were validated. Additional terms were added to the literature search protocol in 2022 to capture data related to drinking and wastewater treatment. When the search string terms are identified in the title, abstract or as a keyword of a given reference in SWIFT-Review, those references proceed with TIAB screening.

3.4 Environmental Release and Occupational Exposure

The searches for peer-reviewed and gray literature are described in Sections 4.2 and 4.3, respectively, in the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Specifically, SWIFT-Review was used to identify peer-reviewed references that are predicted to be the most relevant for evaluating environmental release and occupational exposure for the Risk Evaluation for BBP (U.S. EPA, 2025q). As described in Sections 4.2.4.2 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), EPA identified on-topic and off-topic references from the broad search results of the BBP peer-reviewed literature as positive and negative "seeds" to classify which references contained environmental release and occupational exposure to prioritize for further review. When the relevant references were identified in SWIFT Review, those references proceeded with title and abstract screening.

3.5 General Population, Consumer, and Environmental Exposure

The peer-reviewed and gray literature searches for general population, consumer, and environmental exposure are as described in Sections 4.2 and 4.3, respectively, in the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Specifically, SWIFT-Review was used to identify peer-reviewed references

that are predicted to be the most relevant for evaluating general population, consumer, and environmental exposures to BBP. As described in Sections 4.2.4.2 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), EPA identified on-topic and off-topic references from the broad search results of the peer-reviewed literature as positive and negative "seeds" to classify which references on general population, consumer, and environmental exposures to prioritize for further review. As noted previously in Section 3.1, additional references were added to the literature search protocol to capture database data from NHANES and the Canadian Government database. The database data were compared to other database and monitoring data found during the literature search to ensure no duplication of data. There were no other changes to the process identified in the 2021 Draft Systematic Review Protocol for information considered for the evaluation of general population, consumer, and environmental exposure to BBP (U.S. EPA, 2021).

3.6 Environmental and Human Health Hazard

The search for peer-reviewed and gray literature are as described in Sections 4.2 and 4.3, respectively, in the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Specifically, SWIFT-Review was used to identify peer-reviewed references that are predicted to be the most relevant for evaluating environmental and human health hazard for BBP. Specifically, search strings were developed for the two hazard disciplines by EPA's Office of Research and Development (ORD) in collaboration with SWIFT-Review developer, Sciome. As mentioned above in Section 3.1, the search string used to identify potentially relevant peer-reviewed data references for evaluation of the environmental and human health hazard of BBP were validated. When the search string terms are identified in the title, abstract or as a keyword of a given reference in SWIFT-Review, those references proceed with TIAB screening. The environmental and human health hazard search strings are provided online.

As described in Sections 6.5.1 and 6.5.2, in addition to using data from sources that underwent systematic review, additional data identified during evidence integration that might not have undergone systematic review on BBP were also considered.

3.7 Dermal Absorption

As described above in Section 3.1, EPA used a key word list (search string) to filter the literature identified in the 2019 BBP search to find potentially relevant information for the characterization of dermal absorption of BBP. The search string is listed below (Section 3.7.1).

3.7.1 Dermal Absorption Search String

"Dermal flux" OR "Skin flux" OR "Dermal penetration" OR "Skin penetration" OR "Dermal absorption fraction" OR "Absorption fraction" OR "Neat Kp" OR "Aqueous Kp" OR "Kp" OR "Skin permeability coefficient" OR "Permeability coefficient" OR "Skin permeation coefficient" OR "Permeation coefficient" OR "Skin permeation" OR "Skin absorption" OR "Dermal absorption" OR "Dermal permeation" OR "OECD 427" OR "OECD 428".

4 DATA SCREENING

Sections 4.2.5 and 4.3.2 of the 2021 Draft Systematic Review Protocol describe how TIAB and full-text screening respectively, are conducted to identify references that may contain relevant information for use in risk evaluations under TSCA using discipline-specific screening criteria (U.S. EPA, 2021). Specifically, TIAB screening efforts may be conducted using the specialized web-based software programs DistillerSR[1] and SWIFT-Active-Screener,[2, 3] and the below sub-sections will describe whether TIAB screening was done manually in DistillerSR or utilized machine learning to help prioritize reference screening in SWIFT-Active-Screener. Additional details on how SWIFT Active-Screener utilizes a machine-learning algorithm to automatically compute which unscreened documents are most likely to be relevant[4] are available in Section 4.2.5 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). During TIAB screening, if it was unclear whether a reference met the screening criteria (*e.g.,* PECO/RESO/PESO statements) without having the full reference to review, or if a reference was determined to meet the screening criteria, that reference advanced to full-text screening if the full reference could be retrieved and generated into a Portable Document Format (PDF).

Literature inventory trees were introduced in the scoping process for the risk evaluations that began systematic review in 2019 in response to comments received from the SACC and public to better illustrate how references underwent various systematic review steps (*e.g.,* TIAB and full-text screening). As explained in Section 2.1.2 of the *Final Scope of the Risk Evaluation for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2020b), literature inventory trees demonstrate how references that meet screening criteria progress to the next systematic review step. EPA used the Health Assessment Workplace Collaborative (HAWC) tool to develop web-based literature inventory trees that enhance the transparency of the decisions resulting from the screening processes.

Additional references that were not part of the original 2019 literature search on BBP, but that EPA has obtained via public or other sources (*e.g.,* identified in searches for other chemicals undergoing risk evaluations, chemical assessor identified, backward searches) were also considered in the systematic review process and are reflected in the interactive HAWC hyperlinks available in the figure captions below each respective literature inventory tree. The web-based interactive literature inventory trees in HAWC also allow users to directly access the references in the Health & Environmental Research Online (HERO) database (more details available in Section 1 of the 2021 Draft Systematic Review Protocol). Instructions for accessing information about references and data sources in each node via HERO are available in HAWC for each respective literature inventory tree. Each node indicates whether a reference has met screening criteria at different screening steps and/or contains types of content that may be discerned at that respective systematic review step (U.S. EPA, 2021). Furthermore, the sum of the numbers for the various nodes in the literature inventory trees may be smaller or larger than the

[1] As noted on the DistillerSR web page, this systematic review software "automates the management of literature collection, triage, and assessment using AI and intelligent workflows...to produce transparent, audit ready, and compliant literature reviews." EPA uses DistillerSR to manage the workflow related to screening and evaluating references; the literature search is conducted external to DistillerSR.

[2] SWIFT-Active Screener is another systematic review software that EPA is adopting in the TSCA systematic review process. From Sciome's SWIFT-Active Screener web page: "As screening proceeds, reviewers include or exclude articles while an underlying statistical model in SWIFT-Active Screener automatically computes which of the remaining unscreened documents are most likely to be relevant. This 'Active Learning' model is continuously updated during screening, improving its performance with each reference reviewed. Meanwhile, a separate statistical model estimates the number of relevant articles remaining in the unscreened document list."

[3] SWIFT is an acronym for "*Sciome Workbench for Interactive Computer-Facilitated Text-mining.*" SWIFT-Active Screener uses machine learning approaches to save screeners' time and effort.

[4] Description comes from the SWIFT-Active Screener web page.

preceding node because some studies may have unclear relevance or be relevant for many categories of information. The screening process for each discipline varies and the nodes in the literature inventory tree indicate the screening decisions determined for each reference and whether specific content could be determined; if no references had a specific screening decision and/or contained specific content relevant for a respective discipline, a node will not be present on the literature tree to depict this.

Occasionally some references or data sources are identified in the literature search because of the availability of the title and abstract, however EPA may not be able to always locate the entire or original version. Therefore, references or data sources that meet TIAB screening criteria may be unattainable for full-text screening. The "PDF not available" node within the literature inventory tree refers to references that were identified in the literature search, but which EPA was unable to obtain the entire reference or source of information.

While all information contained in references that enter systematic review is considered for use in the risk evaluation, the references that satisfy the screening criteria are generally deemed to contain the most relevant and useful information for characterizing the uses of, exposure to, and hazard associated with a chemical of interest and are generally utilized in the risk evaluation or to identify further data needs. On the other hand, data or information sources that do not satisfy the screening criteria outlined below may undergo data quality evaluation and extraction should a data need arise for the risk evaluation.

4.1 Multi-disciplinary Updates and Clarifications to the Data Screening

As stated above in Section 1, all references that are found in the initial chemical-specific searches are considered for use in the respective chemical risk evaluation. Previously Section 4.2.5 of the 2021 Draft Systematic Review Protocol explained that references tagged as potentially having supplemental information may be considered for data quality evaluation and extraction. However, one clarification to that description is that even references that are tagged as not meeting TIAB or full-text screening criteria (*e.g.*, PECO/PESO/RESO) for a respective discipline or sub-discipline may also undergo additional screening to meet information needs that were not stated in the original screening criteria and be considered for data quality evaluation and extraction, should there be additional relevant information that may not have met the original screening criteria.

Section 3.1 described that new literature was included in systematic review from additional data sources submitted to the Agency by the SACC during the peer review of the *Draft Risk Evaluation for Butyl Benzyl Phthalate (BBP)* as well as additional references submitted during the public comment period of the risk evaluation of BBP. After an initial review by EPA, the additional data sources were identified to support environmental release and occupational exposure, general population, consumer, and environmental exposures as well as hazard data sources related to environmental toxicity and human health hazard. References from the literature submitted to EPA identified to potentially have information on environmental release and occupational exposure were screened as previously described in Section 4.2.5 and Appendix H.3 of the 2021 Draft Systematic Review Protocol. References from the literature submitted to EPA identified to potentially have information on general population, consumer, and environmental exposures were screened as previously described in Section 4.2.5 and Appendix H.4 of the 2021 Draft Systematic Review Protocol. As described in Section 4.6.1 an updated hazard PECO statement was developed to screen references submitted to EPA by the SACC and during the public comment period of the risk evaluation of BBP and identified as potentially having hazard information related to environmental toxicity and human health hazard. This updated hazard PECO statement was employed to prioritize and narrow down references that were most relevant and filled data gaps.

4.2 Physical and Chemical Properties

During data screening, EPA followed the process described in Appendix H, Section H-1 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), to conduct title and abstract and full-text screening for BBP guided by the data or information needs on various physical and chemical properties or endpoints as listed in Table_Apx H-1 of the protocol. The same screening criteria was used during TIAB and FT screening for references considered for the evaluation of physical and chemical properties of BBP. Title and abstract screening were performed using SWIFT Active-Screener. Upon meeting screening criteria during full-text screening, data or information sources then undergo data quality evaluation and extraction. Figure 4-1 presents the number of references that report general physical and chemical property information that fulfilled the data needs for BBP and passed these criteria for TIAB and FT screening.

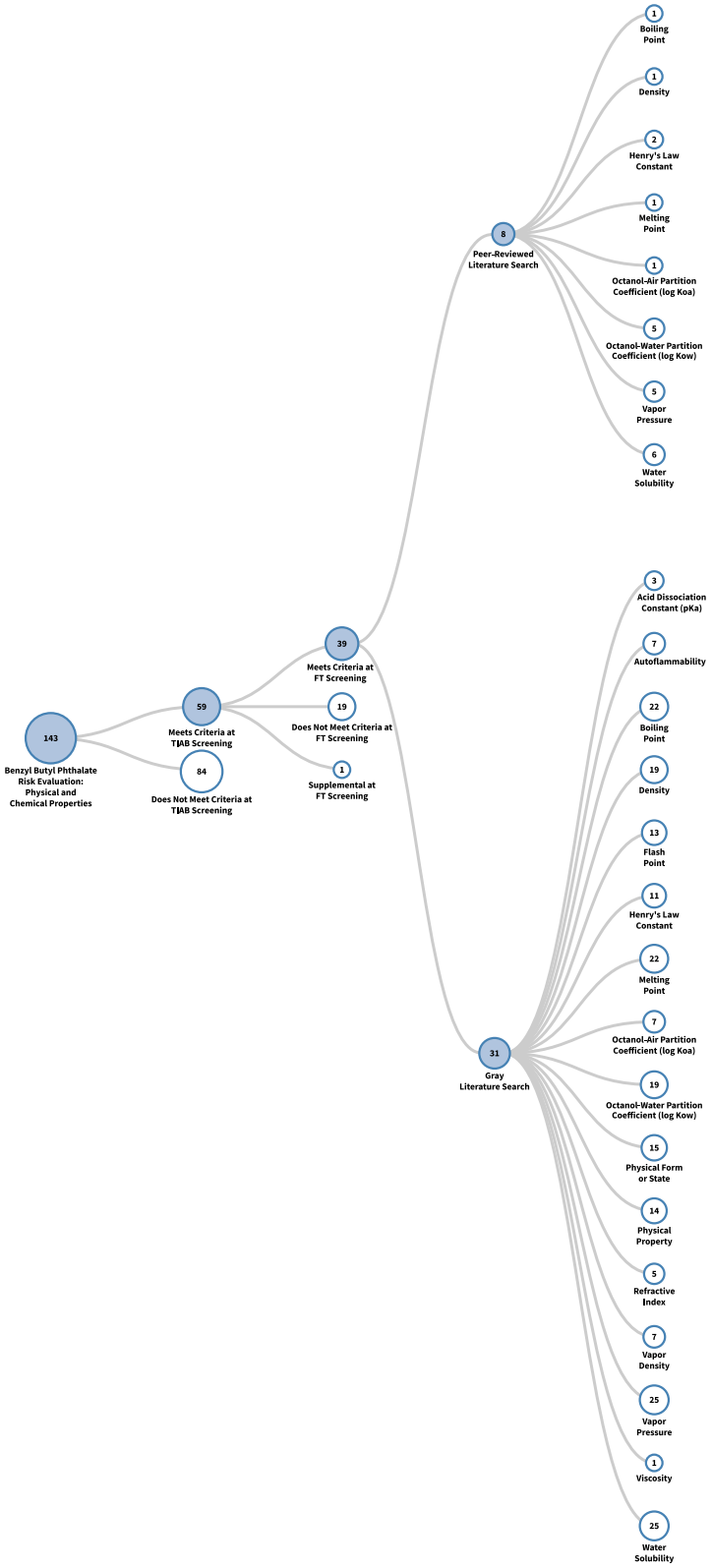


Figure 4-1. Literature Inventory Tree – Physical and Chemical Properties for BBP
View the interactive literature inventory tree in HAWC. Data in this figure represent all references obtained from the publicly available databases and gray literature reference searches that were included in systematic review as of January 31, 2024. Additional data may be added to the interactive version as they become available. Some studies may be found through multiple searches and may have more than one source tag in HERO.

4.3 Environmental Fate and Transport Properties

During data screening, EPA followed the process described in Appendix H, Section H.2 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), to conduct TIAB and FT screening for BBP literature search results, as guided by the PESO statement. PESO stands for **P**athways or **P**rocesses, **E**xposure, **S**etting or **S**cenario, and **O**utcomes (see Table_Apx H2 in 2021 Draft Systematic Review Protocol). The same PESO screening criteria was used during TIAB and FT screening for references considered for the evaluation of environmental fate and transport properties of BBP. TIAB screening was performed using SWIFT Active-Screener. Data or information sources that comply with the screening criteria specified in the PESO statement then undergo data quality evaluation and extraction. Figure 4-2 presents the number of references that report BBP fate processes and endpoints, or environmental and exposure pathways that passed PESO screening criteria at TIAB and FT screening.

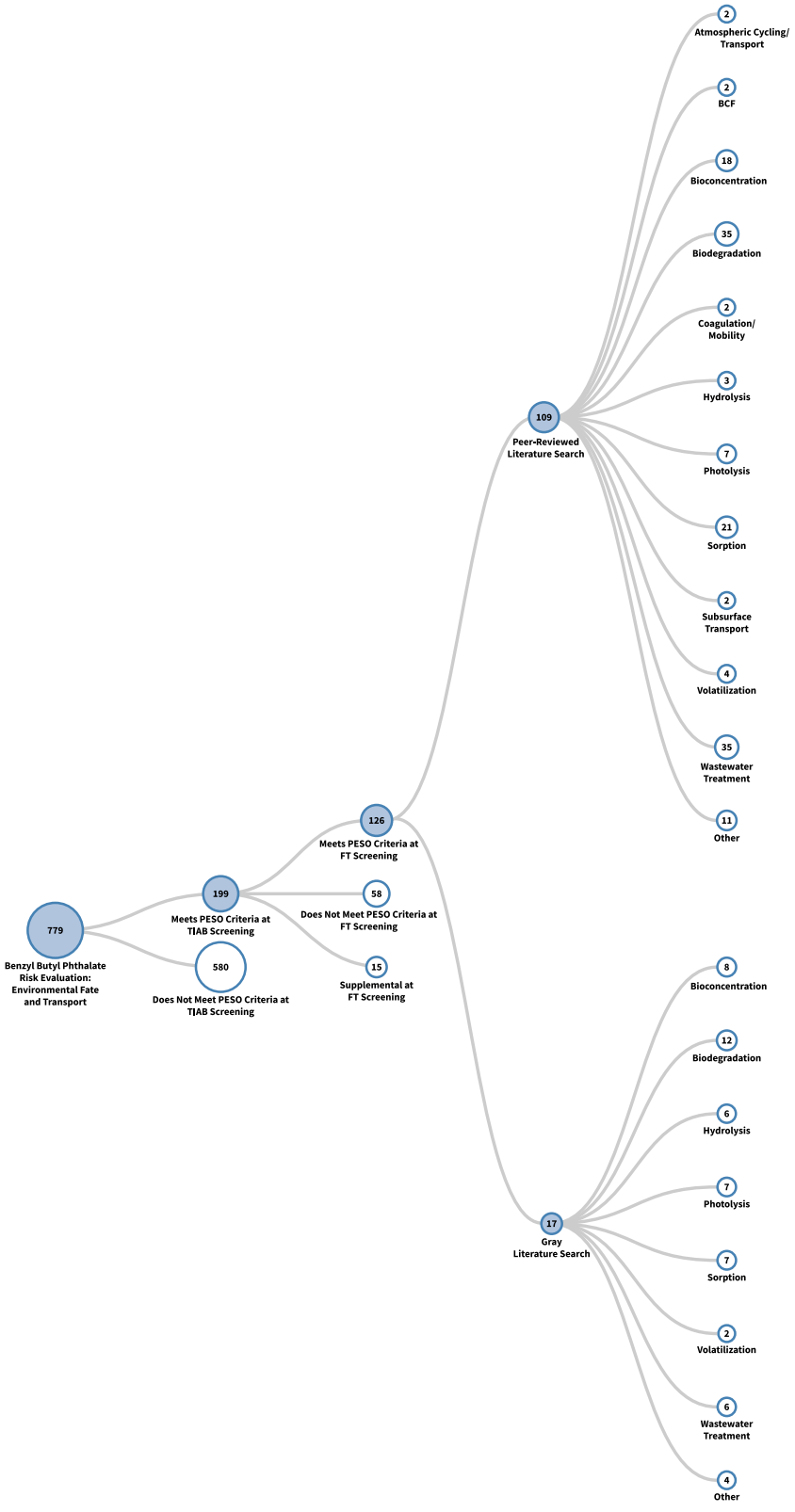


Figure 4-2. Literature Inventory Tree – Environmental Fate and Transport Properties for BBP
View the interactive literature inventory tree in HAWC. Data in this figure represent all references obtained from the publicly available databases and gray literature references searches that were included in systematic review as of January 22, 2025. Additional data may be added to the interactive version as they become available.

4.4 Environmental Release and Occupational Exposure

During data screening, EPA followed the process described in Appendix H, Section H.3 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), to conduct title and abstract, and full-text screening for BBP literature search results, as guided by the RESO statement. RESO stands for **R**eceptors, **E**xposure, **S**etting or Scenario, and **O**utcomes. The same RESO statement was used during title and abstract, and full-text screening for references considered for the evaluation of environmental release and occupational exposure information for BBP. TIAB were performed using SWIFT Active-Screener. Data or information sources that comply with the screening criteria specified in the RESO statement then undergo data quality evaluation and extraction. Figure 4-3 presents the number of references that report general engineering data, environmental release, and occupational exposure data that passed RESO screening criteria at TIAB, and full-text screening.

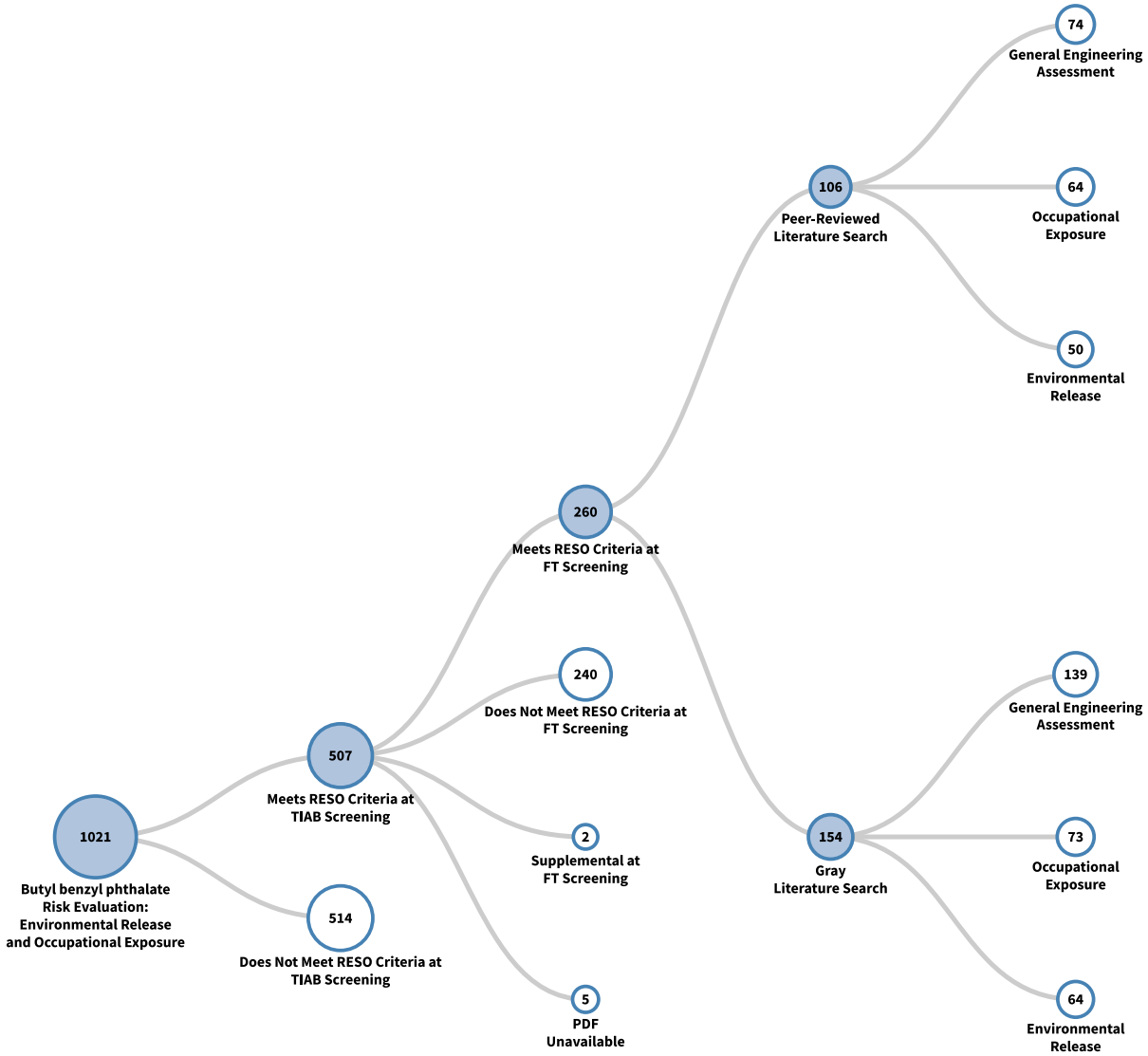


Figure 4-3. Literature Inventory Tree – Environmental Release and Occupational Exposure for BBP

View the interactive literature inventory tree in HAWC. Data in this figure represent all references obtained from the publicly available databases and gray literature references searches that were included in systematic review as of January 27, 2025. Additional data may be added to the interactive version as they become available.

4.5 General Population, Consumer, and Environmental Exposure

During data screening, EPA followed the process described in Appendix H.4 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), to conduct TIAB and full-text screening for BBP literature search results, as guided by the PECO statement. PECO stands for **P**opulation, **E**xposure, **C**omparator or Scenario, and **O**utcomes for Exposure Concentration or Dose. The same PECO statement was used during TIAB and full-text screening for references considered for the evaluation of general population, consumer, and environmental exposure information for BBP. TIAB screening was performed using SWIFT Active-Screener. Figure 4-4 presents the number of references that report general population, consumer, and environmental exposure data that passed PECO screening criteria at TIAB and full-text screening.

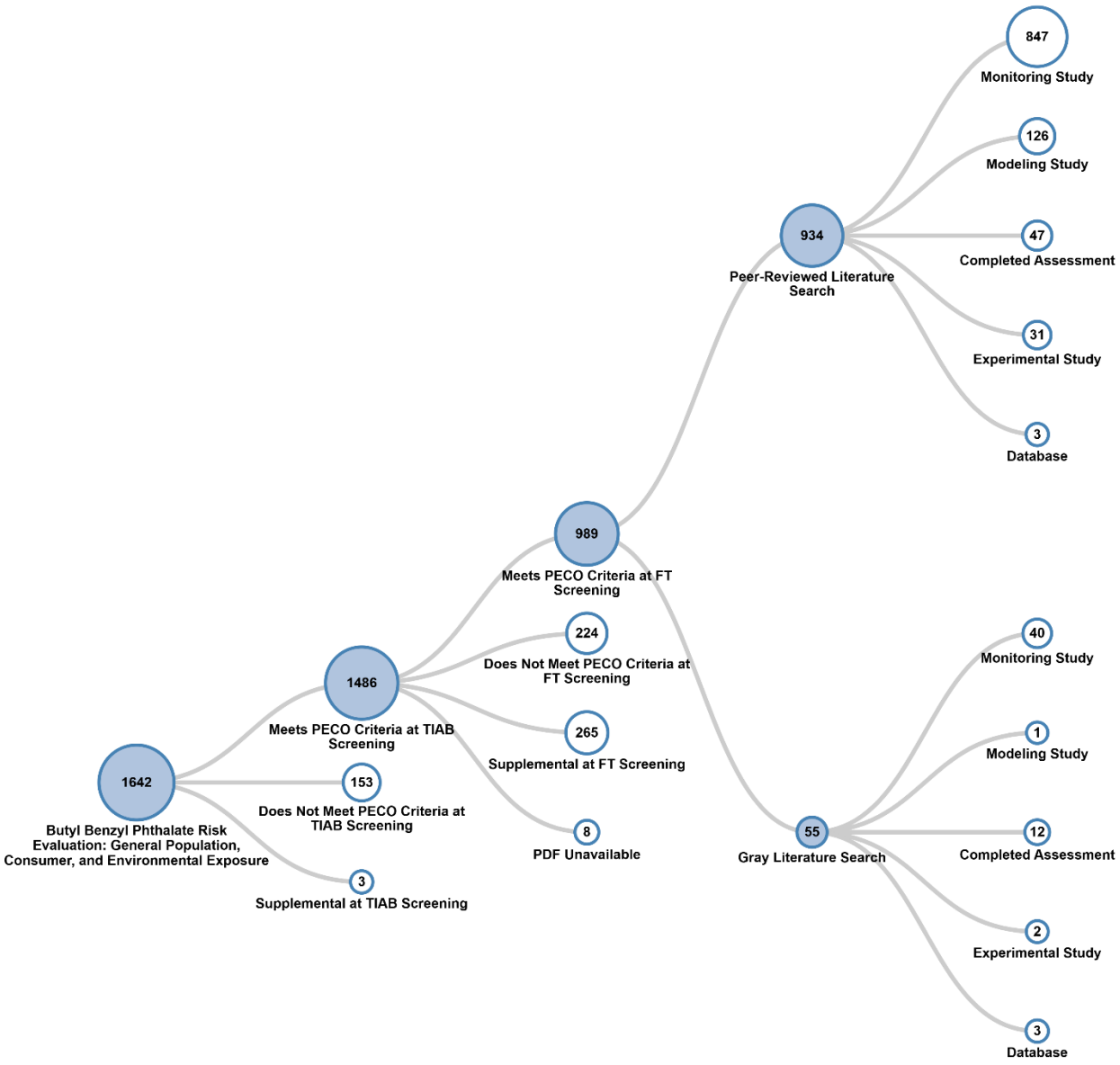


Figure 4-4. Literature Inventory Tree – General Population, Consumer, and Environmental Exposure Search Results for BBP

View the interactive literature inventory tree in HAWC. Data in this figure represent all references obtained from the publicly available databases and gray literature references searches that were included in systematic review as of January 6, 2025. Additional data may be added to the interactive version as they become available.

4.5.1 Further Filtering: General Population, Consumer, and Environmental Exposure

A targeted approach was implemented to the systematic review of BBP references for certain media types based on the priorities and rationales to address key data needs for the exposure assessment Figure 4-4. References that met the PECO screening criteria and were categorized as having exposure information for the evaluation of exposure studies went through a fit-for-purpose further filtering step to determine which studies would move forward to data quality evaluation and data extraction.

As summarized in Section 1 of the *Environmental Media and General Population Exposure for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025m), EPA focused on U.S. studies to compare against EPA's own analysis of NHANES biomonitoring data. BBP concentrations in ambient air, surface water, sediment, soil, landfills, and biosolids were gathered and summarized within each environmental media pathway within the Environmental Exposure Media Concentrations Technical Support Package (U.S. EPA, 2025m). The sources and approaches to gather monitoring data from peer-reviewed publications, government reports, and/or databases were classified as monitoring and mainly used to compare with modeling results or to support qualitative assessments. Consumer products containing BBP were identified through review and searches of a variety of sources, such as completed assessments, 2016 and 2020 Chemical Data Reporting (U.S. EPA, 2020a, 2016). General population and environmental exposures were evaluated for the inhalation, dermal and ingestion exposure pathways based on environmental release data. In summary, modeled environmental release estimates were used as inputs for the general population exposure modeling. To evaluate general population and environmental exposures based on measured and predicted concentrations of BBP in ambient air, reported measured concentrations for ambient air found in the peer-reviewed from the systematic review and the estimated ambient air concentrations from Section 3.1 and 3.2 of the Risk Evaluation for BBP (U.S. EPA, 2025q) were used. To assess environmental exposure, EPA prioritized measured concentrations of BBP within published literature for surface water, precipitation, and sediment.

4.6 Environmental and Human Health Hazard

During data screening, EPA followed the process described in Appendix H, Section H.5.11 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), to conduct TIAB and full-text screening for BBP literature search results, as guided by the PECO statement. In addition to BBP, the PECO statement for phthalates in Appendix H.5.11 also included the various other phthalates that are undergoing a risk evaluation under TSCA: dibutyl phthalate, diethylhexyl phthalate, di-isobutyl phthalate, dicyclohexyl phthalate, diisodecyl phthalate, and diisononyl phthalate. PECO stands for **P**opulation, **E**xposure, **C**omparator or Scenario, and **O**utcomes for Exposure Concentration or Dose. The same PECO statement was used during TIAB and full-text screening for references considered for the evaluation of environmental and human health hazard resulting from exposure to BBP. For TIAB screening, EPA utilized machine learning to help prioritize reference screening in SWIFT-Active-Screener. Full-text screening occurred in DistillerSR for references that either met the PECO screening criteria during TIAB screening or if it was unclear to EPA whether the reference would meet the PECO screening criteria based on the information available in the title and abstract.

Although the PECO statement provided in Appendix H.5.11 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021) was used during TIAB and full-text screening, there is one clarification. Under the **E**xposure PECO element, EPA listed the relevant forms for the various phthalates, including BBP, undergoing a risk evaluation under TSCA. For human (epidemiological) studies, the criteria for the **E**xposure PECO element also included exposure as measured by common metabolites that were described as being specified in a list. However, the list of common metabolites of each phthalate (including BBP) was inadvertently omitted from Appendix H.5.11 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Therefore, listed here is the common metabolites of BBP that EPA

considered during the screening of epidemiological studies: Mono-n-benzyl phthalate (MBzP) and mono-butyl phthalate (MBP).

As described in Sections 3.1 and 4.1, in addition to the sources identified in the original literature search in 2019, EPA identified new literature from additional data sources submitted to the Agency by the SACC during the peer review of the *Draft Risk Evaluation for Butyl Benzyl Phthalate (BBP)* as well as additional references submitted during the public comment period of the risk evaluation of BBP. The PECO statement used to conduct TIAB and full-text screening for the updated literature search was updated from what was published in Appendix H.5.11 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Specific updates to the PECO screening criteria that were used to screen new references are described below in Section 4.6.1.

On July 10, 2024, EPA received supplemental information from B&C® Consortia Management, L.L.C. (BCCM). The consortia members submitted data sources related to ecotoxicity data supporting the risk evaluation for BBP. Between the draft and final risk evaluation of BBP, EPA has considered all data sources submitted and identified new data sources. These new data sources were screened for relevancy during TIAB and full-text screening. The PECO statement used to conduct TIAB and full-text screening is described in Section 4.6.1.

Figure 4-5 presents the number of references that report environmental and human health hazard data that met PECO screening criteria at TIAB and full-text screening for BBP.

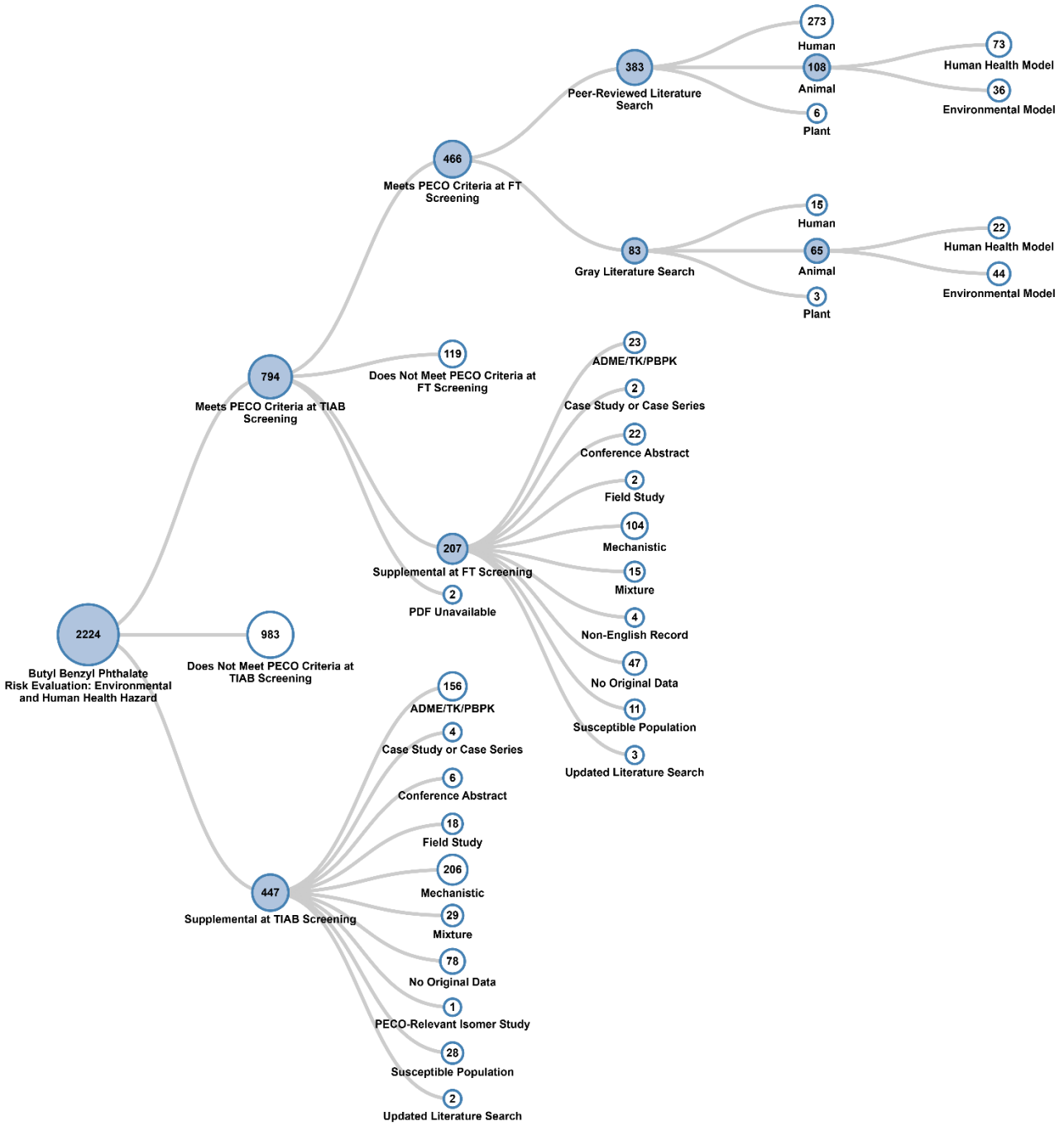


Figure 4-5. Literature Inventory Tree – Environmental and Human Health Hazard for BBP
View the interactive literature inventory tree in HAWC. Data in this figure represent all references obtained from the publicly available databases and gray literature references searches that were included in systematic review as of November 19, 2025. Additional data may be added to the interactive version as they become available.

4.6.1 Hazard Targeted PECO Screening Criteria Updates

As part of the new literature update, to screen references submitted to EPA by the SACC during the peer review of the *Draft Risk Evaluation for Butyl Benzyl Phthalate (BBP)* as well as references submitted during the public comment period of the risk evaluation of BBP in addition to references submitted by BCCM, EPA updated the PECO statement for BBP (Table 4-1 and Table 4-2). The screening criteria were developed as a targeted approach to prioritize the information that was most relevant and presented new information for characterizing both environmental and human health hazard for the risk evaluation for BBP. Because sometimes references reporting information on the target phthalate (*i.e.*, BBP) also reported information on other phthalates, the updated PECO statement reflects how information reported on BBP as well as other phthalates undergoing a risk evaluation under TSCA was screened. To make it easier for the reader to see changes made or clarifications added to the screening criteria published in Appendix H.5.11 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), the following conventions are used in Table 4-1 and Table 4-2 text inserted is underlined, and text deleted is in strikethrough.

The targeted approach to prioritize new information that was most relevant for human health hazard did not include epidemiological studies, and therefore the updated PECO statement in Table 4-3 does not include any screening criteria for human epidemiological studies (this is indicated with N/A in Table 4-1). EPA did not prioritize new human epidemiological studies because this information would not have had a quantitative impact on the human health hazard assessment, because the body of epidemiological evidence indicated that humans are exposed to multiple phthalates from multiple sources via multiple routes, resulting in substantial uncertainties in quantifying exposure-response estimates for individual phthalates. Therefore, new human epidemiological information from an updated literature search would not have influenced the human health assessment in a quantitative manner.

Because the updated PECO statement in Table 4-1 includes updated criteria based on exposure route, exposure duration, dosages used, and outcome (*e.g.*, cancer and non-cancer) to help identify new information that would fill data gaps, animal toxicity studies informing human health hazard were not further filtered with the Further Filtering Form described in Table 4-3 which was utilized for references from the 2019 literature search (Figure 4-6, Box 1a). Therefore, any reference from the new literature update that met the updated PECO criteria was subsequently evaluated using the Harmonized TSCA Data Quality Evaluation Form in Table 5-7 and data were extracted as described in Section 5.5.1.

Table 4-1. Updated PECO Criteria for: Butyl benzyl phthalate (CASRN 85-68-7), Dibutyl phthalate (CASRN 84-74-2), Di(2-ethylhexyl) phthalate (CASRN 117-81-7), Diisobutyl phthalate (CASRN 84-69-5), and Dicyclohexyl phthalate (CASRN 84-61-7) – Title and Abstract and Full-Text Screening

PECO Element	Clarification
Population	**Human:** ~~Any population and lifestage (*e.g.*, occupational or general population, including children and other sensitive populations~~. N/A. (Studies on humans were not considered for systematic review of the 2025 new literature update. Human studies at this time for the 2025 new literature update will be tagged as ***Supplemental, Updated literature search: Meets original PECO criteria but does not fill a critical data gap***.) **Animal:** Aquatic and terrestrial species (live, whole organism) from any lifestage (*e.g.*, preconception, in utero, lactation, peripubertal, and adult stages). Animal models will be inventoried according to the categorization below: • *Human health models*: rat, mouse, rabbit, dog, hamster, guinea pig, cat, non-human primate, pig, ~~hen (neurotox only)~~. • *Ecotoxicological models*: ~~invertebrates (*e.g.*, insects, spiders, crustaceans, mollusks, and worms) and vertebrates (*e.g.*, mammals and all amphibians, birds, fish, and reptiles)~~. All animal studies (invertebrates and vertebrates) excluding the models listed above as a human health model. All hen studies (including neurotoxicity studies) will meet PECO screening criteria as ecotoxicological animal models. **Plant:** All aquatic and terrestrial species (live) (vascular and non-vascular plants), including but not limited to algal species, diatoms, cyanobacteria, moss, lichen and macro fungi (*e.g.*, mushrooms (Phylum: Basidiomycota)) species. **Screener notes:** • Human Health Animal Hazard and Environmental Hazard: To identify human health and ecological hazards, other organisms not listed above in their respective categories can also be used. Non-mammalian model systems are increasingly used to identify potential human health hazards (*e.g., Xenopus*, zebrafish), and traditional human health models (*e.g.*, rodents) can be used to identify potential ecological hazard. For systematic review screening and data evaluation and extraction purposes, the human health models listed above will be tagged or identified as human health models and all other animal studies will be tagged as ecotoxicological animal models. Neurotoxicity studies performed in hens (*e.g.*, OECD 418 and 419) are considered relevant to both human health and environmental hazard, but all hen studies will be tagged only as ecotoxicological animal models for systematic review screening and data evaluation and extraction purposes. • Environmental Hazard: Ecotoxicological studies that assess exposure effects on organisms such as protozoan, microbial fungi (*e.g.*, microsporidians) and molds do not meet PECO screening criteria because

PECO Element	Clarification
	an environmental hazard assessment will unlikely be driven by unicellular organisms or microbial organisms which are low in the natural ecosystem hierarchy. • Environmental Hazard: The Population (PECO) consideration should be directed toward direct effects on the **target species only** regardless of the type of effect or health outcome. Studies reporting only indirect effects expressed in taxa that are **not** the target species of the chemical exposure do **not** meet the PECO screening criteria. • Human Health Animal Hazard and Environmental Hazard: Studies on gametes, embryos, or plant (*e.g.*, ungerminated seeds, harvested fruit, cut flowers, and potato tubers) or fungal sections capable of forming whole, new organisms will be tagged as potentially ***Supplemental, Mechanistic***. EXCEPTION: For environmental hazard, embryos for animal studies (*e.g.*, zebrafish, fathead minnow, copepod, bivalve embryos, chickens) and germinated seeds for plant studies (*e.g.*, seed germination in any plant) meet screening criteria if they also meet all other PECO criteria. • Human Health Animal Hazard and Environmental Hazard: Bacteria and yeast studies specific for assessing genotoxicity, mutagenicity (*e.g.*, Ames assay), or hormone assay will be tagged as potentially ***Supplemental, Mechanistic***. Otherwise, bacteria and yeast studies that are not used for assessing genotoxicity, mutagenicity, or hormone assays do **not** meet the PECO criteria. • Human Health Animal Hazard and Environmental Hazard: Studies on viruses and any pathogenic microbes (unless bacteria or yeast used for assessing genotoxicity, mutagenicity, or hormone assay; see bullet above) do **not** meet the PECO screening criteria.
Exposure	**Relevant forms:** • Butyl benzyl phthalate (BBP) (CASRN 85-68-7) • Dibutyl phthalate (DBP) (CASRN 84-74-2) • Diethylhexyl phthalate (DEHP) (CASRN 117-81-7) • ***Isomer***: Isooctyl phthalate - 27554-26-3 • Di-isobutyl phthalate (DIBP) (CASRN 84-69-5) • Dicyclohexyl phthalate (DCHP) (CASRN 84-61-7) • For synonyms see the EPA Chemistry Dashboard. **Human:** ~~Any exposure singularly or in mixture, including exposure as measured by internal concentrations of these chemicals or metabolites of these chemicals in a biological matrix (*i.e.*, urine, blood, semen, etc.). See list of common metabolites for each phthalate below~~. N/A. (Studies on humans were not considered for systematic review of the 2025 new literature update. Human studies at this time for the 2025 new literature update will be tagged as ***Supplemental, Updated literature search: Meets original PECO criteria but does not fill a critical data gap***.)

PECO Element	Clarification
	Animal Human Health Models: Any exposure to BBP, DBP, DEHP, DIBP, and/or DCHP including via water (including environmental aquatic exposures), soil or sediment, diet, gavage, injection, dermal, and inhalation for all non-rodent species. In order to target data gaps identified in previous literature searches, rodent exposure should be limited to 1) inhalation at all doses, 2) dermal at all doses, and 3) oral studies evaluating cancer at all doses, and 4) oral studies with a non-cancer effect at or below the point of departure (POD) for BBP, DBP, DEHP, DIBP, DCHP used in the draft risk evaluation of the target chemical. When it is difficult to determine effect level, studies with at least one dose at or below the POD should be included for DBP, BBP, DEHP, DIBP, and DCHP. Exposure routes in rodent studies beyond inhalation, dermal and oral as specified in the previous statements will be tagged as ***Supplemental, Updated literature search: Meets original PECO criteria but does not fill a critical data gap***. **Animal Ecotoxicological Models:** Any exposure to BBP, DBP, DEHP, DIBP, and/or DCHP including via water (including environmental aquatic exposures), soil or sediment, diet, gavage, injection, dermal, and inhalation. **Plant:** Any exposure to BBP, DBP, DEHP, DIBP, and/or DCHP including via water or soil, or sediment. EXCEPTION for Environmental Hazard: Waterborne studies with exposure concentrations above the limit of water solubility will be tagged ***Supplemental, Updated literature search: Meets original PECO criteria but does not fill a critical data gap***. **Screener notes:** • Environmental Hazard: Field studies with media concentrations (*e.g.*, surface water soil sediment) and/or body/tissue concentrations of animals or plants are to be identified as ***Supplemental, Field Study*** only if any biological effects are reported. • Environmental Hazard: Controlled outdoor experimental studies (*e.g.*, controlled crop/greenhouse studies, mesocosm studies, artificial stream studies) are considered to be laboratory studies (**not field studies**) because there is a known and prescribed exposure dose(s) and an evaluation of hazardous effect(s). Whereas field studies (*e.g.*, biomonitoring) where there is no prescribed exposure dose(s) do **not** meet the PECO screening criteria if there is no evaluated hazardous effect, and tagged as ***Supplemental, Field study*** if there is an evaluated hazardous effect. • Human Health Animal Hazard and Environmental Hazard: Studies involving exposures to mixtures will be included only if they also include exposure to BBP, DBP, DEHP, DIBP, or DCHP alone. Otherwise, ~~mixture studies will be tagged as Supplemental~~ mixture studies in human health animal models will be tagged as ***Supplemental, Mixture study (human***

PECO Element	Clarification
	health animal models) and mixture studies in ecotoxicological animal models and in plants will be tagged as ***Supplemental, Mixture study (plants and eco health animal models).***
Comparator	**Human:** ~~A comparison or referent population exposed to lower levels (or no exposure/exposure below detection limits) of BBP or exposure to BBP for shorter periods of time.~~ N/A. (Studies on humans were not considered for systematic review of the 2025 new literature update. Human studies at this time for the 2025 new literature update will be tagged as ***Supplemental, Updated literature search: Meets original PECO criteria but does not fill a critical data gap***.) **Animal and Plants:** A concurrent control group exposed to vehicle-only treatment and/or untreated control (control could be a baseline measurement). **Screener notes:** • Human Health Animal Hazard and Environmental Hazard: If no control group is explicitly stated (implied (*e.g.*, by mention of statistical results that could only be obtained if a control group was present), the study will be marked as ***Unclear*** during **Title/Abstract Screening.** During **Full-text Screening**, if no control group is explicitly stated, then the study does **not** meet PECO screening criteria. • Human Health Animal Hazard: For studies in which human health animal models are intentionally exposed to a chemical, the control could be a baseline measurement of the same individual (*i.e.*, the individual is assessed pre- and post-exposure), and these studies do meet the PECO screening criteria. Also, for studies in which human health animal models are intentionally exposed to a chemical, references that contain experimental designs that do not require a negative or vehicle control group (*i.e.*, skin sensitization (such as LLNA), LC50 and LD50 completed within an acute timeframe, or dermal irritation studies in which the experimental individual serves as their own control) do meet the PECO screening criteria.
Outcome	**Human:** ~~All health outcomes (cancer and noncancer) at the organ level or higher~~. N/A. (Studies on humans were not considered for systematic review of the 2025 new literature update. Human studies at this time for the 2025 new literature update will be tagged as ***Supplemental, Updated literature search: Meets original PECO criteria but does not fill a critical data gap***.) **Animal Human Health Models**: All apical biological effects (effects measured at the organ level or higher) and bioaccumulation from laboratory studies with concurrently measured media and/or tissue concentrations for exposure routes of interest. Apical endpoints include but are not limited to reproduction, survival, and growth. All studies evaluating cancer should be included regardless of dose, route, or species. Oral studies in rodents should be limited to 1) evaluation of cancer and 2) a non-cancer effect at or

PECO Element	Clarification
	below the point of departure (POD) for DBP, BBP, DEHP, DIBP, DCHP used in the risk evaluation. When it is difficult to determine effect level, studies with at least one dose at or below the POD should be included for DBP, BBP, DEHP, DIBP, and DCHP. Oral studies in rodents that report non-cancer health outcomes above the POD used in the draft risk evaluation will be tagged as ***Supplemental, Updated literature search: Meets original PECO criteria but does not fill a critical data gap***. **Animal Ecotoxicological Models and Plants:** All health outcomes in *in vivo* studies including mechanistic studies reporting exposure to DIBP and DCHP. Mortality, growth, development, and reproductive outcomes will be prioritized for DBP, BBP, and DEHP. Studies that do not report these health outcomes for DBP, BBP, and DEHP but would have otherwise met the original PECO from 2019 will be tagged as ***Supplemental, Updated literature search: Meets original PECO criteria but does not fill a critical data gap***. **Screener notes:** • Environmental Hazard: If the study has definitive hazard data for at least one health outcome (bounded values) as reported by authors, the study is prioritized for data evaluation and extraction. However, if the study that is being screened only has non-definitive hazard data (unbounded values for all reported health outcomes) as reported by authors, the study will not be prioritized for data evaluation and extraction and will be tagged as ***Supplemental, Updated literature search: Meets original PECO criteria but does not fill a critical data gap.*** • Environmental Hazard: For DBP, BBP, and DEHP, chronic aquatic studies will only be prioritized if the hazard value reported by authors is below the hazard value used to calculate the concentration of concern (COC) in the draft risk evaluation of the phthalate of interest. For DBP, BBP, and DEHP, if the chronic aquatic studies only report hazard values above the hazard value used to calculate the concentration of concern (COC) in the draft risk evaluation of the phthalate of interest, the study will be tagged as ***Supplemental, Updated literature search: Meets original PECO criteria but does not fill a critical data gap.*** • Human Health Animal Hazard and Environmental Hazard: Measurable biological effects relevant for animals and plants may include but are not limited to: mortality, behavioral, population, physiological, growth, reproduction, systemic, point of contact (irritation and sensitization) effects. • Human Health Animal Hazard and Environmental Hazard: Effects measured at the cellular level of biological organization and below are to be tagged as ***Supplemental, Mechanistic.***

Table 4-2. Major categories of Potentially Relevant Supplemental Material for: Butyl benzyl phthalate (CASRN 85-68-7), Dibutyl phthalate (CASRN 84-74-2), Di(2-ethylhexyl) phthalate (CASRN 117-81-7), Diisobutyl phthalate (CASRN 84-69-5), and Dicyclohexyl phthalate

Category	Evidence
Mechanistic studies	All studies that report results at the cellular level and lower in both mammalian and non-mammalian model systems, including *in vitro*, *in vivo*, *ex vivo*, and *in silico* studies. These studies include assays for genotoxicity or mutagenicity using bacteria or yeast.
ADME, PBPK, and toxicokinetic	Studies designed to capture information regarding absorption, distribution, metabolism, and excretion (ADME), toxicokinetic studies, or physiologically based pharmacokinetic (PBPK) models.
Field studies	Field studies with media concentrations (e.g., surface water, interstitial water, soil, sediment) and/or body/tissue concentrations of animals or plants if biological effects reported.
Mixture studies	Experimental mixture studies that are not considered PECO-relevant because they do not contain an exposure or treatment group assessing only the chemical of interest. Human health animal model and eco animal model/plant will be tagged separately for mixture studies.
Non-English records	Non-English records will be tracked as potentially relevant supplemental information.
Records with no original data	Records that do not contain original data, such as other agency assessments, informative scientific literature reviews, editorials or commentaries, but may cite secondary data on dermal absorption. This also includes studies of dermal exposure, risk, or modeling that may cite dermal absorption studies.
Conference abstracts	Records that would otherwise meet PECO criteria, but do not contain sufficient documentation to support study evaluation and data extraction.
Updated literature search: Meets original PECO criteria but does not fill a critical data gap	Studies that met the original PECO as published in the *Draft Systematic Review Protocol Supporting TSCA Risk Evaluations for Chemical Substances* (U.S. EPA, 2021), however, they did not fill critical data gaps as per the additional criteria described in the revised PECO statement.

4.6.2 Further Filtering: Human Health Hazard

References that met the PECO screening criteria and were categorized as having epidemiology information and/or animal toxicity information for the evaluation of human health hazard went through a fit-for-purpose further filtering step to determine which studies would move forward to data quality evaluation and data extraction.

4.6.2.1 Epidemiology Studies

To streamline the identification of studies containing dose-response data that had not previously been evaluated by EPA, modifications were implemented to the process described in the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Following PECO-based screening, references that met PECO screening criteria for epidemiology underwent a two-step further filtering process to identify the subset of potentially relevant references that proceeded to data quality evaluation:

4.6.2.1.1 Epidemiology Further Filtering Step 1: Filtering for References Published After the Literature Search End Date of the Most Recent Authoritative Assessment

The first step of further filtering consisted of filtering for references published after the literature search end date of the most recent authoritative assessment. Previous phthalates risk assessments have been conducted by authoritative sources including Health Canada and the EPA IRIS program. OPPT used these previous assessments to facilitate efficient and scientific risk evaluation. Therefore, data quality evaluation and extraction were conducted only for references published after the literature search end date of the most recent authoritative assessment.

The most recent authoritative assessment was published by Health Canada in 2020 and included literature published up to 2018 (Health Canada, 2020). Therefore, data quality evaluation and extraction were conducted for references published from the beginning of 2018 through the end date of the OPPT literature search, as well as for references that were published from the beginning of 2018 through the end of 2023 that were sent with public comments in phthalates dockets. Data quality evaluation and extraction wasn't conducted for any references published before 2018.

Previous assessments used phthalates epidemiology studies qualitatively, but epidemiology studies weren't used quantitatively for dose-response assessment. Therefore, no key studies were identified from previous assessments. Furthermore, all BBP references may be of interest qualitatively. Therefore, further filtering wasn't used to identify or filter for dose-response studies.

Thus, the first step of further filtering was based only on publication date. Labels were added in DistillerSR to indicate references with publication dates of 2018 or later. All BBP references that met PECO screening criteria for epidemiology with a publication date of 2018 or later proceeded to the next step of further filtering. All other BBP references (references with a publication date before 2018) didn't proceed to data quality evaluation.

4.6.2.1.2 Epidemiology Further Filtering Step 2: Filtering Out References That Only Assessed Exposure Using an Inappropriate Biomarker Matrix

Urine is generally the only appropriate biomarker matrix for assessing exposure to short-chain phthalates and primary metabolites of long-chain phthalates. The IRIS Protocol for the Systematic Review of the Health Effects of Phthalate Exposure describes the reasons why biomarker matrices other than urine are inappropriate for assessing exposure. The IRIS Protocol states "Phthalate metabolite concentration in urine is considered to be the best proxy of exposure from all sources (ingested/absorbed/inhaled). One of the problems with phthalates measured in blood and other tissues is the potential for contamination from outside sources, especially during the collection and processing of samples (Calafat et al., 2015). Phthalate diesters present from exogenous contamination can be metabolized to the monoester metabolites by enzymes present in blood and other tissues (but not urine). Thus, metabolite measures in samples other than urine may be erroneously reflecting external phthalate sources" (Radke et al., 2020; Radke et al., 2018).

Therefore, in the IRIS phthalates assessment, "biomarker measures based on samples other than urine (*e.g.*, serum, plasma, amniotic fluid, seminal fluid, amniotic fluid, breast milk) were considered to be critically deficient for all short-chain phthalates and for primary metabolites (*e.g.*, MEHP, MINP) of long-chain phthalates" (Radke et al., 2020; Radke et al., 2018). Although breast milk is not an appropriate biomarker matrix for assessing the exposure of the person who produced the milk, phthalate measures from breast milk are appropriate for assessing exposure to infants who are ingesting the breast milk.

The IRIS protocol states "Samples other than urine can be used for secondary metabolites of long-chain phthalates as the oxidative metabolism required to break down primary metabolites does not exist in these samples (personal communication, Antonia Calafat, 2016). Cord blood, as a sample matrix, is considered critically deficient for all metabolites, since DEHP (and possibly BBP) containing plastics are widely used in medical settings, and thus, the concentrations of phthalates in cord blood may reflect exposure during delivery. In addition, studies that analyzed only phthalate diesters, rather than their metabolites, are considered critically deficient due to the potential for contamination" (Radke et al., 2020; Radke et al., 2018).

Therefore, data quality evaluation wasn't conducted for references that assessed exposure using *only* a biomarker matrix other than urine or breast milk without any other exposure assessment. Otherwise, all epidemiology references that met PECO screening criteria, had a publication date of 2018 or later, and used a potentially appropriate exposure assessment method proceeded to data quality evaluation.

4.6.2.1.3 Epidemiology Further Filtering Results

Of the 291 peer-reviewed references that met BBP PECO screening criteria for epidemiology, step 1 of the further filtering process identified 110 references that had a publication date of 2018 or later, which proceeded to step 2 of the further filtering process. Out of these 110 references, 6 references were found to assess exposure using only non-urine biomarkers and therefore didn't proceed to data quality evaluation. The remaining 104 references proceeded to data quality evaluation for BBP.

4.6.2.2 Animal Toxicity Studies

Studies that met the PECO screening criteria and were categorized as having animal toxicity information for the evaluation of human health hazard were then identified to either have been previously evaluated by an authoritative agency or not. Studies that had not previously been evaluated by an authoritative agency and were considered relevant for animal toxicity went through a more extensive further filtering process similar to that described in the previous section (4.6.2.1) to identify and prioritize animal toxicity studies with quantitative information most useful for the human health hazard assessment.

4.6.2.2.1 Animal Toxicity Further Filtering Step 1: Identification of Whether or Not Studies were Cited in a Recent Authoritative Assessment

During full-text screening, 95 studies were identified to meet the PECO screening criteria for animal toxicity informing human health hazard (Figure 4-6, Box 1). Previous phthalates risk assessments have been conducted by authoritative sources including Health Canada (EC/HC) (EC/HC, 2015b). OPPT used this previous assessment to facilitate an efficient and scientific risk evaluation. Based on this existing assessment, a total of 10 key studies were considered for point of departure (POD) refinement (Ahmad et al., 2014; Kwack et al., 2009; Howdeshell et al., 2008; Lee and Koo, 2007; Tyl et al., 2004; Nagao et al., 2000; TNO CIVO, 1993; BIBRA, 1986). Thus, these 10 studies did not go through a further filtering step and moved directly to the data evaluation and extraction step under TSCA (Figure 4-6, Box 2a and 3a). References that underwent further filtering were those published after the EC/HC 2015 assessment up until the literature search conducted by OPPT for the BBP risk evaluation, which covered the years

2014 – 2019. One reference was added by assessors to aide in meta-analysis during POD refinement (Gray et al., 2021). No new studies were added to the reference pool following the 2025 Updated Literature search which included references from Public Comment and EPA SACC review period discussed in Section 3.1.

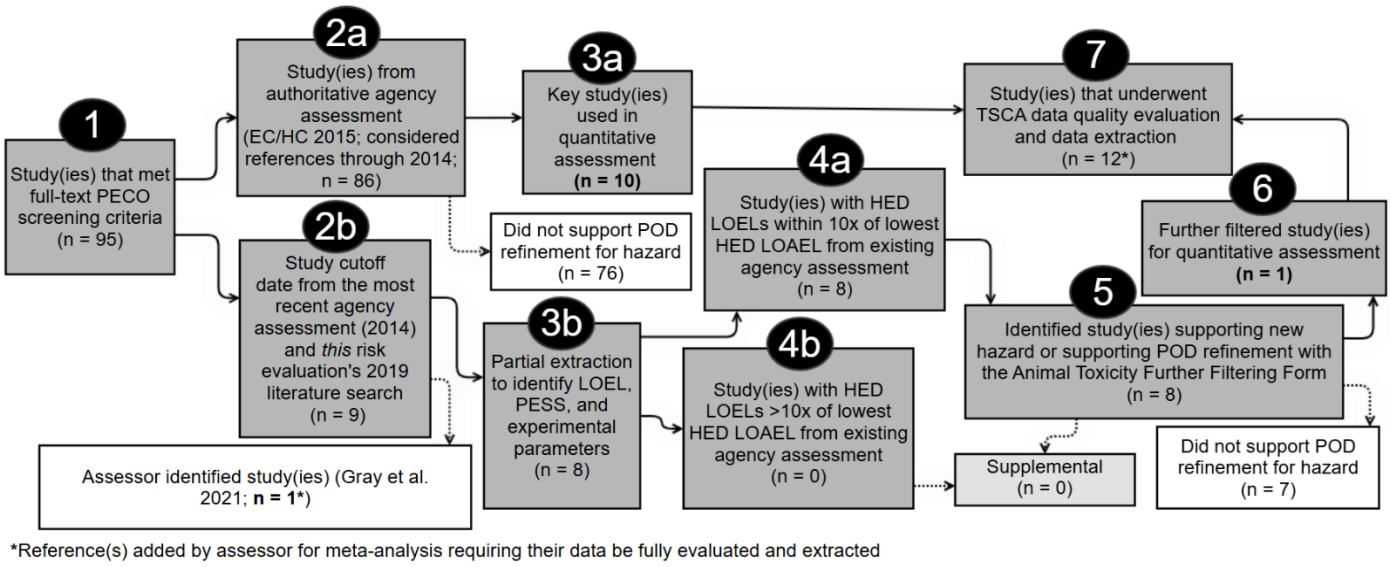


Figure 4-6. Schematic for the Number of Animal Toxicity Information for Human Health Hazard for BBP that were Evaluated and Extracted under TSCA

4.6.2.2.2 Animal Toxicity Further Filtering Step 2: Identification of Studies Used in EPA's Quantitative Assessment

For the remaining 8 studies that were published after the EC/HC 2015 assessment, study parameters and lowest-observable-effect levels (LOELs) were then collected (Figure 4-6, Box 3b) and converted to human equivalent doses (HEDs) to enable comparisons across species. Studies with HED LOELs greater than an order of magnitude of the lowest HED lowest-observable-adverse-effect level (LOAEL) identified across existing assessments were not deemed sensitive for subsequent POD selection (Figure 4-6, Box 4b). For BBP, there were no studies that fell in this category (Figure 4-6, Box 4b); however, if any studies had been identified, they would have been tagged as supplemental information, and they would not have proceeded to data quality evaluation and extraction.

On the contrary, studies with HED LOELs within an order of magnitude of the lowest HED LOAEL identified across existing assessments were considered sensitive and potentially relevant for POD selection (Figure 4-6, Box 4a). For BBP, there were 8 such studies identified and were further reviewed by EPA to determine if they provided information that either supported a new human health hazard not identified in the existing assessments, or to determine if the 8 studies contained sufficient dose-response information to support a lower POD than identified in the existing assessments (Figure 4-6, Box 5). Next, these 8 studies (Jahreis et al., 2018; ILS, 2017; Nakagomi et al., 2017; Debartolo et al., 2016; Schmitt et al., 2016; Ahmad et al., 2015; Alam and Kurohmaru, 2015; Min et al., 2014) were filtered using the Animal Toxicity Further Filtering form described in Table 4-3. The Animal Toxicity Furter Filter Form was developed to tag and identify studies by exposure route, exposure method and duration of exposure, number of dose groups, target organ/systems evaluated, information related to potentially exposed or susceptible subpopulations (PESS), and the study-wide LOEL. The main purpose of this

further filtering step was to allow for the refinement of the references that would be considered for data quality evaluation and extraction. For BBP, of the 8 studies that went through the Animal Toxicity Further Filtering Form, only 1 study moved on to data quality evaluation and extraction (Ahmad et al., 2015) (Figure 4-6, Box 6), while the remaining were no longer considered for POD refinement. In section 4.6.2.2.3, EPA describes in detail the decisions made for studies that went through the Animal Toxicity Further Filtering Form.

4.6.2.2.3 Further Filtering Results

Out of 8 remaining studies that went through the Animal Toxicity Further Filtering Form (Figure 4-6, Box 5), EPA determined that Ahmad et al. (2015) was the only reference that proceeded to full data quality evaluation and extraction. At the end, a total of 12 animal toxicity studies for the data integration of human health hazard were evaluated and extracted for BBP under TSCA (Figure 4-6, Box 7). For a detailed list of health outcomes and ratings along with a description and rationale for such ratings as well as details on which data were extracted, see the *Risk Evaluation for Butyl Benzyl Phthalate (BBP) – Systematic Review Supplemental File: Data Quality Evaluation Information for Human Health Hazard Animal Toxicology* (U.S. EPA, 2025j) and the *Risk Evaluation for Butyl Benzyl Phthalate (BBP) – Systematic Review Supplemental File: Data Extraction Information for Environmental Hazard and Human Health Hazard Animal Toxicology and Epidemiology* (U.S. EPA, 2025b).

Table 4-3. Animal Toxicity Further Filtering Form

Animal Toxicity Further Filtering Form	
Is this study a candidate for re-screening? (*i.e.*, PECO-relevance related issues) If yes, please stop inventorying?	
○ Yes ▪ Reason for re-screen [free text]	
○ No	
Animal Species	
○ Cat	○ Pig
○ Dog	○ Primate
○ Guinea Pig	○ Rabbit
○ Hamster	○ Rat
○ Mouse	○ Other ○ [free text]
Exposure Routes (check all that apply)	
○ Inhalation	○ Ocular/Eye
○ Dermal/Skin	○ Intraamniotically
○ Oral ▪ If 'Yes' ▫ Gavage, Drinking Water, Food, or Capsule	○ Other ○ Other exposure routes (describe) [free text]
○ Injection	

Animal Toxicity Further Filtering Form	
▪ If 'Yes' ▫ Intraperitoneal ▫ Subcutaneous	
Is this a reproductive/developmental study?	
o Yes	
o No	
Select Study Duration:	
o Acute (≤ 24 hr)	
o Short-Term (> 1-30 days)	
o Sub-Chronic (> 30 – 90 days)	
o Chronic (> 90 days)	
o Not Reported	
Does this study contain 2 or more dose groups in addition to a control?	
o Yes	
o No	
Please inventory target organs/systems with outcomes reported (qualitative or quantitative, including negative outcomes):	
o Neurological/Behavioral	o Immune/Hematological
o Cancer/Carcinogenesis	o Hepatic/Liver
o Cardiovascular	o Mortality
o Thyroid	o Musculoskeletal
o Reproductive/Developmental	o Nutritional/Metabolic
o Gastrointestinal	o Ocular/Sensory
Does this study report a LOEL?	
o Yes ▪ Experiment LOEL dose value [free text] ▪ LOEL Units (mg/kg-bw/day, mg/kg, etc.)	
o No	
o Other o [free text]	
o Briefly describe the LOEL outcome [free text]	
Does this study report any negative outcomes (*i.e.*, no change seen in animals following exposure)?	
o Yes	
o No	
Does the experiment show different effects among GENETICS/EPIGENETICS PESS subpopulations (genetic variants that increase susceptibility, knockout animals, etc.)?	
o Yes	
o No	

Animal Toxicity Further Filtering Form
Does the experiment show different effects among LIFESTAGE PESS subpopulations (reproductive studies, accumulation in milk, etc.)?
o Yes
o No
Does the experiment show different effects among OTHER (not listed) PESS subpopulations (reproductive studies, accumulation in milk, etc.)? If so, please list below.
o [free text]
Should this reference move on to data extraction and evaluation?
o Yes
o No
Comments (optional)?
o [free text]

4.7 Dermal Absorption

EPA developed a PECO statement (Table 4-4) to conduct both TIAB and full-text screening of references considered for the evaluation of dermal absorption resulting from BBP exposure. EPA used categories in Table 4-5 to identify supplemental studies that may also inform dermal absorption and exposure for BBP. Each reference was manually screened by two reviewers at the TIAB and full-text screening steps or only at full-text, as relevant for the type of data source (peer vs. gray). Figure 4-7 presents the outcome of applying the search strings presented in Section 3.7.1 and the PECO screening criteria below.

Table 4-4. PECO Statement for Dermal Exposure References for BBP

PECO Element	Evidence
P	Tests of the single toxicants on *ex vivo* tissues (including permeation and retention studies) or on live, whole, taxonomically verifiable organisms are included. **Human:** Any population and life stage (occupational or general population, including children and other sensitive populations). **Animal:** All human health models, including (but not limited to) rat, mouse, rabbit, dog, hamster, guinea pig, cat, non-human primate, and pig. **Supplemental:** Tests using 3D human skin equivalent/reconstructed tissue models (*e.g.*, EpiDerm, EPISKIN) or any other *in vitro* systems are considered supplemental.
E	**Human and Animal:** Any quantified dermal exposure to butyl benzyl phthalate (CASRN 85-68-7) or in a vehicle or relevant matrix associated with the conditions of use, including exposure that occurs *in vivo* or *ex vivo* for any duration. Studies are included only if exposure is intentional and quantified. If exposure is not intentional and is not experimentally controlled, the study is excluded. For example, studies of absorption in workers will be excluded, even if exposure has been quantified. Studies assessing exposures to mixtures (*i.e.*, containing substances other than a vehicle) will be included only if they also contain an exposure or treatment group assessing the chemical of interest alone or in aqueous solution.

PECO Element	Evidence
	Supplemental: *In vitro* exposures and/or studies in which exposure occurs only to a mixture containing one or more of the chemicals of interest.
C	**Human and Animal:** Any or no comparison group.
O	**Human and Animal**: Any quantitative assessment of the rate or extent of dermal absorption of the substance. Measurements may include the amount of substance that has passed through the skin, or was retained in the skin, distributed within the organism (*e.g.*, blood and tissue concentrations), and/or excreted by the organism (*e.g.*, through urine, feces, or expired air). Absorption may be measured directly (by chemical analysis for the substance and/or its metabolites) or indirectly (*e.g.*, measurement of radioactivity if using a radio-labelled test substance). Absorption may be quantified via determination of percent absorption, dermal/penetrative flux rate, or dermal penetration coefficient (Kp).

Table 4-5. Major Categories of "Potentially Relevant Supplemental Material"

Category	Evidence
***In vitro* studies**	Tests using 3D human skin equivalent/reconstructed tissue models (*e.g.*, EpiDerm, EPISKIN) or any other *in vitro* systems.
Mixture studies	Experimental mixture studies that are not considered PECO-relevant because they do not contain an exposure or treatment group assessing only the chemical of interest, but that otherwise meet PECO criteria.
Non-English records	Non-English records that appear to meet PECO criteria.
Records with no original data	Records that do not contain original data, such as other agency assessments, informative scientific literature reviews, editorials, or commentaries, but may cite secondary data on dermal absorption. This also includes studies of dermal exposure, risk, or modeling that may cite dermal absorption studies.
Conference abstracts	Records that would otherwise meet PECO criteria, but do not contain sufficient documentation to support study evaluation and data extraction.

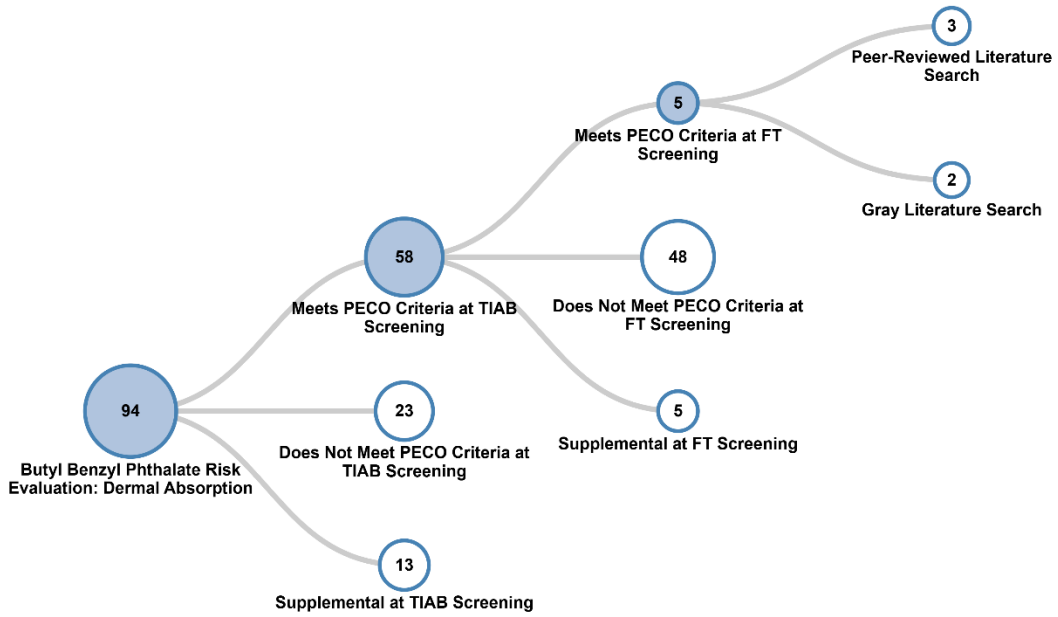


Figure 4-7. Literature Inventory Tree – Dermal Absorption for BBP
View the interactive literature inventory tree in HAWC. Data in this figure represent all references obtained from the publicly available databases and gray literature references searches that were included in systematic review as of November 21, 2025. Additional data may be added to the interactive version as they become available.

5 DATA EVALUATION AND DATA EXTRACTION

Data evaluation and extraction were conducted as described in Sections 5 and 6 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Data evaluation is the systematic review step in which EPA assesses quality of the individual data sources using the evaluation strategies and criteria for each discipline (*e.g.*, physical and chemical property data; fate and transport data; occupational exposure and environmental release data; general population, consumer, and environmental exposure data; environmental hazard; human health hazard) or sub-discipline (*e.g.*, animal toxicity or epidemiology). The data quality evaluation method uses a structured framework with predefined criteria for each type of data/information source. Data extraction is the systematic review step in which EPA uses structured forms or templates to extract quantitative and qualitative data and information from references that meet screening criteria. The overall goal is to provide transparency, consistency, and as much objectivity as possible to the data quality evaluation and extraction processes along with meeting the TSCA scientific standards in section 26(h).

References that meet screening criteria following full-text screening will generally proceed to data quality evaluation and extraction steps, however one clarification to the procedures outlined in Section 6 of the 2021 Draft Systematic Review Protocol is that in situations where EPA is unable to extract data/information from sources that meet screening criteria (*e.g.*, formatting prohibits accurate extraction), that source may not have extracted data to present in the risk evaluation or respective supplemental documents. The systematic review supplemental files that contain results from the data quality evaluation and extraction systematic review steps may use updated templates from those that were provided in the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021) because the purpose of these supplemental documents is to accommodate the data needs for each respective risk evaluation. The following sections describe the data quality and extraction process followed by each discipline or sub-discipline to address various information needs for the Risk Evaluation for BBP (U.S. EPA, 2025q), and any clarifications or updates regarding these systematic review steps as described in the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021).

5.1 Physical and Chemical Properties

As described in the 2021 Draft Systematic Review Protocol, evaluation and extraction followed the steps outlined in Sections 5, 6, and 6.1 (U.S. EPA, 2021). The data quality criteria for physical and chemical property data are summarized in Appendix K of the 2021 Draft Systematic Review Protocol. The *Data Quality Evaluation and Data Extraction Information for Physical and Chemical Properties for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025g) provides details of the data extracted and evaluated, including metric ratings and the overall study quality determination for each data source.

5.2 Environmental Fate and Transport Properties

As described in the 2021 Draft Systematic Review Protocol, evaluation and extraction followed the steps outlined in Sections 5, 6, and 6.2 (U.S. EPA, 2021). The data quality criteria for environmental fate data are summarized in Appendix L of the systematic review protocol. Appendix L.4 describes how the overall quality of fate data or information were weighted according to an ordinal system corresponding to *High (1)*, *Medium (2)*, or *Low (3)* to quantitatively or qualitatively support the risk evaluations. EPA does not plan to use data rated as *Uninformative (4)*. Table_Apx L4 illustrates the possible quality rankings across the selected metrics for environmental fate data with examples in Table_Apx L5, Table_Apx L6 and Table_Apx L7 (U.S. EPA, 2021). Specific fate data quality ranking quality criteria are in Table_Apx L8. The *Data Quality Evaluation and Data Extraction Information for Environmental Fate and Transport for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025e) provides details of the data

extracted and evaluated, including metric rating and the overall study quality determination for each data source.

5.3 Environmental Release and Occupation Exposure

As described in the 2021 Draft Systematic Review Protocol, evaluation and extraction followed the steps outlined in Sections 5, 6, and 6.2 (U.S. EPA, 2021). The data quality criteria for environmental release and occupational exposure data are summarized in Appendix M of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). The *Data Quality Evaluation and Data Extraction Information for Environmental Release and Occupational Exposure for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025f) details the data extracted and evaluated, including metric rating and the overall study quality determination for each data source.

5.4 General Population, Consumer, and Environmental Exposure

As described in the 2021 Draft Systematic Review Protocol, data quality evaluation and extraction generally followed the steps outlined in Section 5 and 6 (U.S. EPA, 2021). However, a few updates were made to the data quality evaluation metrics for some evidence streams (*i.e.*, study types) since the metrics were published in the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Most of the changes were editorial or minor clarifications, including the standardization of some metrics that apply to multiple evidence streams, where appropriate. For example, in the quality assurance/quality control (QA/QC) metric for evaluating monitoring and experimental evidence streams, the acronym QA/QC was defined and replaced all references to quality assurance and quality control when occurring separately or together, and the term "QA/QC techniques" was changed to "QA/QC measures," which already appeared in the metrics.

A few metrics applicable to multiple evidence streams were slightly modified to better fit some of the unique situations that frequently arise for a certain type of evidence stream (*e.g.*, databases). For example, some metrics were updated to clarify the intent of the metric and better account for variation in types of evidence included in one grouping (*e.g.*, experiments involving chamber studies vs. product concentration assessments). The domains did not change, however see below for the changes and updates made to the data evaluation metrics for the respective evidence types (*i.e.*, monitoring, experimental studies and databases) as presented in Section 5.4.1. No changes were made to the data evaluation metrics for modeling data, as described in Appendix N.6.2, or to the data evaluation metrics for completed exposure assessments and risk characterizations, as described in Appendix N.6.7 in the 2021 Draft Systematic Review Protocol, respectively (U.S. EPA, 2021). Data quality evaluations for references that met PECO screening criteria are included in the *Data Quality Evaluation Information for General Population, Consumer, and Environmental Exposure for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025i), referred to hereafter as the "BBP Data Quality Evaluation Information for General Population, Consumer, and Environmental Exposure."

Data extraction of general population, consumer, and environmental exposure data and information was conducted as described in Section 6 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). However, with respect to information stored within databases, if EPA has access to the data tables, EPA does not conduct a separate data extraction because the data are more accessible and have additional context in the original database format. Data present in the database when the database underwent full-text screening are available in the HERO database (CDC, 2022; U.S. EPA, 2022a, b; QuanTech, 2021), along with the date the data were downloaded. If a reference (*e.g.*, peer-reviewed reference) presents data from a database that did not undergo systematic review directly (*e.g.*, a foreign database that is not publicly accessible), the data would be extracted from the reference to the extent possible; this did not apply to references that underwent systematic review for this chemical.

As mentioned above in Section 5, references may not undergo data extraction, regardless of the overall quality determination, if they contain no extractable data points (*e.g.*, values are contained in a non-digitizable figure or are representative of unspecified media or treatment processes). On the other hand, there are references that have many reported endpoints that meet PECO screening criteria for a respective chemical risk evaluation, making it difficult to include all the data in the chemical-specific data extraction supplemental file. When a reference meets PECO screening criteria, the reference receives a data quality evaluation, and the data in the reference are still considered in the Risk Evaluation, whether or not the included data are extracted in DistillerSR and appear among the chemical-specific extractions in the Systematic Review Supplemental File: Data Extraction Information for General Population, Consumer, and Environmental Exposure. In addition, there may be other reasons that EPA decides not to extract all the data from a reference that undergoes data evaluation; EPA extracts the data that are most relevant, given the needs of the assessment. As seen in Figure 4-5, the extracted BBP data are from targeted evaluated references that have an OQD of High assuming that such studies would be distinctly supportive to the BBP exposure assessment. The extracted data provide a high level of confidence for characterizing general population, consumer, and environmental exposure and for meeting assessment needs. This constitutes an update to Section 6 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Extraction forms, templates, and decisions are tailored to fit the data extraction needs for each risk evaluation.

The types of fields extracted vary by evidence stream and generally followed Section 6.3 of the 2021 Draft Systematic Review Protocol with regard to the data characteristics captured (U.S. EPA, 2021). Examples of types of data extracted and the extraction formats for the evidence streams identified through systematic review to evaluate environmental, general population, and consumer exposure data are listed in the extraction tables provided in the *Data Extraction Information for General Population, Consumer, and Environmental Exposure for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025c), referred to hereafter as the "BBP Data Extraction Information for General Population, Consumer, and Environmental Exposure."

5.4.1 Data Quality Evaluation Metric Updates

The data evaluation metrics for the monitoring, experimental, and database evidence streams, are presented below in Table 5-1, Table 5-2, and Table 5-3, respectively. Each table shows which data evaluation metrics changed since the publication of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Other data quality criteria for studies on consumer, general population, and environmental exposure appear in Appendix N of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). For the modeling, completed exposure assessments, and risk characterization evidence streams, there were no changes made to the data evaluation metrics since the 2021 Draft Systematic Review Protocol was published. The criteria for modeling studies appear in Table_Apx N-9 of the 2021 Draft Systematic Review Protocol, and criteria for completed exposure assessments and risk characterizations appear in Table_Apx N-19. In some cases, references can meet the criteria for two exposure evidence streams, and they can also be reviewed and meet criteria for other disciplines. Upon review, each study is evaluated and extracted using the criteria for the most appropriate and applicable evidence streams given the information therein. In order to make it easier for the reader to see changes made to the data evaluation metrics, the following conventions are used: text inserted is underlined, and text deleted is in ~~strikethrough~~.

Table 5-1. Updated Data Quality Evaluation Criteria for Monitoring Data Sources

Data Quality Rating	Description
Domain 1. Reliability	
Metric 1. Sampling methodology	
High	Samples were collected according to publicly available SOPs that are scientifically sound and widely accepted (*i.e.,* from a source generally ~~using~~ known to use sound methods and/or approaches) for the chemical and media of interest. Example SOPs include U.S. Geological Survey (USGS') "National Field Manual for the Collection of Water-Quality Data," EPA's "Ambient Air Sampling" (SESDPROC-303-R5), etc. OR The sampling protocol used was not a publicly available SOP from a source generally known to use ~~using~~ sound methods and/or approaches, but the sampling methodology is clear, appropriate (*i.e.,* scientifically sound), and similar to widely accepted protocols for the chemical and media of interest. All pertinent sampling information is provided in the data source or companion source. Examples include: • sampling equipment • sampling procedures/regimen • sample storage conditions/duration • performance/calibration of sampler • study site characteristics • matrix characteristics
Medium	Sampling methodology is discussed in the data source or companion source and is generally appropriate (*i.e.*, scientifically sound) for the chemical and media of interest; however, **one or more pieces of sampling information is not described.** The missing information is unlikely to have a substantial impact on results. OR Standards, methods, protocols, or test guidelines may not be widely accepted, but a successful validation study for the new/unconventional procedure was conducted prior to the sampling event and is consistent with sound scientific theory and/or accepted approaches. Or a review of information indicates the methodology is acceptable and differences in methods are not expected to lead to lower quality data.
Low	Sampling methodology is only briefly discussed; therefore, **most sampling information is missing** and likely to have a substantial impact on results. AND/OR The sampling methodology **does not represent best sampling methods, protocols, or guidelines** for the chemical and media of interest (*e.g.*, outdated [but still valid] sampling equipment or procedures, long storage durations). AND/OR There are **some inconsistencies** in the reporting of sampling information (*e.g.*, differences between text and tables in data source, differences between standard method and actual procedures reported to have been used, etc.) that led to a low confidence in the sampling methodology used.
Critically Deficient	The sampling methodology is not discussed in the data source or companion source. AND/OR

Data Quality Rating	Description
	Sampling methodology is not scientifically sound or is not consistent with widely accepted methods/approaches for the chemical and media being analyzed (*e.g.*, inappropriate sampling equipment, improper storage conditions). AND/OR There are **numerous inconsistencies** in the reporting of sampling information, resulting in high uncertainty in the sampling methods used.
Not rated/not applicable	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 2. Analytical methodology	
High	Samples were analyzed according to publicly available analytical methods that are scientifically sound and widely accepted (*i.e.*, from a source generally ~~using~~ known to use sound methods and/or approaches) and are appropriate for the chemical and media of interest. Examples include EPA SW-846 Methods, NIOSH Manual of Analytical Methods 5th Edition, etc. OR The analytical method used was not a publicly available method from a source generally ~~using~~ known to use sound methods and/or approaches, but the methodology is clear and appropriate (*i.e.*, scientifically sound) and similar to widely accepted protocols for the chemical and media of interest. All pertinent sampling information is provided in the data source or companion source. Examples include: • extraction method • analytical instrumentation (required) • instrument calibration • limit of quantitation (LOQ), LOD, detection limits, and/or reporting limits • recovery samples • biomarker used (if applicable) • matrix-adjustment method (*i.e.*, creatinine, lipid, moisture)
Medium	Analytical methodology is discussed in detail and is clear and appropriate (*i.e.*, scientifically sound) for the chemical and media of interest; however, **one or more pieces of analytical information is not described**. The missing information is unlikely to have a substantial impact on results. AND/OR The analytical **method may not be standard/widely accepted, but a method validation study was conducted** prior to sample analysis and is expected to be consistent with sound scientific theory and/or accepted approaches. AND/OR Samples were collected at a site and immediately analyzed using an on-site mobile laboratory, rather than shipped to a stationary laboratory.
Low	Analytical methodology is only briefly discussed. Analytical instrumentation is provided and consistent with accepted analytical instrumentation/methods. However, **most analytical information is missing** and likely to have a substantial impact on results.

Data Quality Rating	Description
	AND/OR Analytical method **is not s**tandard/widely accepted, and method validation is limited or not available. AND/OR Samples were analyzed using field screening techniques. AND/OR LOQ, LOD, detection limits, and/or reporting limits not reported. AND/OR There are **some inconsistencies or possible errors** in the reporting of analytical information (*e.g.*, differences between text and tables in data source, differences between standard method and actual procedures reported to have been used, etc.) which leads to a lower confidence in the method used.
Critically Deficient	Analytical methodology is not described, **including analytical instrumentation** (*i.e.*, HPLC, GC). AND/OR Analytical methodology is not scientifically appropriate for the chemical and media being analyzed (*e.g.*, method not sensitive enough, not specific to the chemical, out of date). AND/OR There are numerous inconsistencies in the reporting of analytical information, resulting in high uncertainty in the analytical methods used.
Not rated/ Not applicable	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 3. Selection of biomarker of exposure	
High	Biomarker in a specified matrix is known to have an accurate and precise quantitative relationship with external exposure, internal dose, or target dose (*e.g.*, previous studies (or the current study) have indicated the biomarker of interest reflects external exposures). AND Biomarker (parent chemical or metabolite) is derived from exposure to the chemical of interest.
Medium	Biomarker in a specified matrix has accurate and precise quantitative relationship with external exposure, internal dose, or target dose. AND Biomarker is derived from multiple parent chemicals, not only the chemical of interest, **but** there is a stated method to apportion the estimate to only the chemical of interest

Data Quality Rating	Description
Low	Biomarker in a specified matrix has accurate and precise quantitative relationship with external exposure, internal dose, or target dose. AND Biomarker is derived from multiple parent chemicals, not only the chemical of interest, **and** there is NOT an accurate method to apportion the estimate to only the chemical of interest. OR Biomarker in a specified matrix is a poor surrogate (low accuracy and precision) for exposure/dose.
Critically Deficient	Not applicable. A study will not be deemed critically deficient based on the use of biomarker of exposure.
Not rated/ applicable	Metric is not applicable to the data source.
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 2. Representative	
Metric 4. Geographic area	
High	Geographic location(s) is reported, discussed, or referenced.
Medium	Not applicable. This metric is dichotomous (*i.e.*, high vs. critically deficient).
Low	Not applicable. This metric is dichotomous (*i.e.*, high vs. critically deficient).
Critically Deficient	Geographic location is not reported, discussed, or referenced.
Not rated/ not applicable	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 5. Temporality	
High	Timing of sample collection for monitoring data is consistent with current or recent exposures **(within 5 years)** may be expected.
Medium	Timing of sample collection for monitoring data **is less consistent** with current or recent exposures **(>5 to 15 years)** may be expected.
Low	Timing of sample collection for monitoring data is not consistent with when current exposures **(>15 years old)** may be expected and likely to have a substantial impact on results.
Critically Deficient	Timing of sample collection for monitoring data is **not reported, discussed, or referenced.**

Data Quality Rating	Description
Not rated/ Not applicable	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 6. Spatial and temporal variability	
High	Sampling approach accurately captures variability of environmental contamination in population/scenario/media of interest based on the heterogeneity/homogeneity and dynamic/static state of the environmental system. For example: • **Large sample size** (*i.e.*, ≥10 or more samples for a single scenario). • Use of replicate samples. • Use of systematic or continuous monitoring methods. • Sampling over a sufficient period of time to characterize trends. • For urine, 24-hour samples are collected (vs. first morning voids or spot). • For biomonitoring studies, the timing of sample collected is appropriate based on chemical properties (*e.g.*, half-life), the pharmacokinetics of the chemical (*e.g.*, rate of uptake and elimination), and when the exposure event occurred.
Medium	Sampling approach likely captures variability of environmental contamination in population/scenario/media of interest based on the heterogeneity/homogeneity and dynamic/static state of the environmental system. Some uncertainty may exist, but it is unlikely to have a substantial impact on results. For example: • **Moderate sample size** (*i.e.*, 5–10 samples for a single scenario), or • Use of judgmental (non-statistical) sampling approach, or • No replicate samples. • For urine, first morning voids or pooled spot samples.
Low	Sampling approach poorly captures variability of environmental contamination in population/scenario/media of interest. For example: • **Small sample size** (*i.e.*, <5 samples), or • Use of haphazard sampling approach, or • No replicate samples, or • Grab or spot samples in single space or time, or • Random sampling that does not include all periods of time or locations, or • For urine, un-pooled spot samples.
Critically Deficient	Sample size is not reported. **Single sample** collected per data set. For biomonitoring studies, the timing of sample collected is not appropriate based on chemical properties (*e.g.*, half-life), the pharmacokinetics of the chemical (*e.g.*, rate of uptake and elimination), and when the exposure event occurred.
Not rated/not applicable	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

Data Quality Rating	Description
Metric 7. Exposure scenario	
High	The data closely represent relevant exposure scenario (*i.e.*, the population/scenario/media of interest). Examples include: • amount and type of chemical/product used • source of exposure • method of application or by-stander exposure • use of exposure controls • microenvironment (location, time, climate)
Medium	The data likely represent the relevant exposure scenario (*i.e.*, population/scenario/media of interest). **One or more key pieces of information may not be described** but the deficiencies are unlikely to have a substantial impact on the characterization of the exposure scenario. AND/OR If surrogate data, activities seem similar to the activities within scope.
Low	The data lack multiple key pieces of information, and the deficiencies are likely to have a substantial impact on the characterization of the exposure scenario. AND/OR There are **some inconsistencies or possible errors** in the reporting of scenario information (*e.g.*, differences between text and tables in data source, differences between standard method and actual procedures reported to have been used, etc.) which leads to a lower confidence in the scenario assessed. AND/OR If surrogate data, activities have lesser similarity but are still potentially applicable to the activities within scope.
Critically Deficient	If reported, the exposure scenario discussed in the monitored study does not represent the exposure scenario of interest for the chemical.
Not rated/ Not applicable	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 3. Accessibility/clarity	
Metric 8. Reporting of results	
High	Supplementary or raw data (*i.e.*, individual data points) are reported, allowing summary statistics to be calculated or reproduced. AND Summary statistics are detailed and complete. Example parameters include: • Description of data set summarized (*i.e.*, location, population, dates, etc.) • Range of concentrations or percentiles • Number of samples in data set • Frequency of detection • Measure of variation (coefficient of variation [CV], standard deviation)

Data Quality Rating	Description
	• Measure of central tendency (mean, geometric mean, median) • Test for outliers (if applicable) AND Both adjusted and unadjusted results are provided (*i.e.*, correction for void completeness in urine biomonitoring, whole-volume or lipid adjusted for blood biomonitoring, wet or dry weight for environmental tissue samples or soil samples) [only if applicable].
Medium	Supplementary or raw data (*i.e.*, individual data points) are not reported, and therefore summary statistics cannot be reproduced. AND/OR Summary statistics are reported but are missing one or more parameters (see description for high). AND/OR Only adjusted or unadjusted results are provided, but not both [only if applicable].
Low	Supplementary data are not provided, and summary statistics are missing most parameters (see description for high). AND/OR There are some inconsistencies or errors in the results reported, resulting in low confidence in the results reported (*e.g.*, differences between text and tables in data source, less appropriate statistical methods).
Critically Deficient	There are numerous inconsistencies or errors in the calculation and/or reporting of results, resulting in highly uncertain reported results.
Not Rated/ Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 9. Quality assurance	
High	The study quality assurance/quality control (QA/QC) measures and all pertinent ~~quality assurance~~ QA/QC information is provided in the data source or companion source. Examples include: • Field, laboratory, and/or storage recoveries. • Field and laboratory control samples. • Baseline (pre-exposure) samples. • Biomarker stability • Completeness of sample (*i.e.*, creatinine, specific gravity, osmolality for urine samples) AND No QA/QC ~~quality control~~ issues were identified, or any identified issues were minor and adequately addressed (*i.e.*, correction for low recoveries, correction for completeness).
Medium	The study applied and documented ~~quality assurance/quality control~~ QA/QC measures; however, **one or more pieces of QA/QC information is not described.** Missing information is unlikely to have a substantial impact on results.

Data Quality Rating	Description
	AND No QA/QC ~~quality control~~ issues were identified, or any identified issues were minor and addressed (*i.e.*, correction for low recoveries, correction for completeness).
Low	QA/QC measures ~~Quality assurance/quality control techniques~~ and results were not directly discussed but are implied through the study's use of standard field and laboratory protocols. AND/OR Deficiencies were noted in ~~quality assurance/quality control~~ QA/QC measures that are likely to have a substantial impact on results. AND/OR There are some inconsistencies in the ~~quality assurance~~ QA/QC measures reported, resulting in low confidence in the QA/QC ~~quality assurance/control~~ measures taken and results (*e.g.*, differences between text and tables in data source).
Critically Deficient	QA/QC issues have been identified which significantly interfere with the overall reliability of the study.
Not Rated/ Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 4. Variability and uncertainty	
Metric 10. Variability and uncertainty	
High	The study characterizes variability in the population/media studied. AND Key uncertainties, limitations, and data gaps have been identified. AND The uncertainties are minimal and have been characterized.
Medium	The study has limited characterization of variability in the population/media studied. AND/OR The study has limited discussion of key uncertainties, limitations, and data gaps. AND/OR Multiple uncertainties have been identified but are unlikely to have a substantial impact on results.
Low	The characterization of variability is absent. AND/OR Key uncertainties, limitations, and data gaps are not discussed. AND/OR Uncertainties identified may have a substantial impact on the exposure the exposure assessment
Critically Deficient	Estimates are highly uncertain based on characterization of variability and uncertainty.
Not Rated/	

Data Quality Rating	Description
Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

Table 5-2. Updated Evaluation Criteria for Experimental Data Sources

Data Quality Rating	Metric Description
Domain 1. Reliability	
Metric 1. Sampling Methodology and Conditions	
High	Samples were collected according to publicly available SOPs, methods, protocols, or test guidelines that are scientifically sound and widely accepted from a source generally known to use sound methods and/or approaches such as EPA, NIST, American Society for Testing and Materials, ISO, and ACGIH. OR The sampling protocol used was not a publicly available SOP from a source generally known to use sound methods and/or approaches, but the sampling methodology is clear, appropriate (*i.e.*, scientifically sound), and similar to widely accepted protocols for the chemical and media of interest. All pertinent sampling information is provided in the data source or companion source. Examples include: • sampling conditions (*e.g.*, temperature, humidity) • sampling equipment and procedures • sample storage conditions/duration • performance/calibration of sampler
Medium	Sampling methodology is discussed in the data source or companion source and is generally appropriate (*i.e.*, scientifically sound) for the chemical and media of interest, however, one or more pieces of sampling information is not described. The missing information is unlikely to have a substantial impact on results. OR Standards, methods, protocols, or test guidelines may not be widely accepted, but a successful validation study for the new/unconventional procedure was conducted prior to the sampling event and is consistent with sound scientific theory and/or accepted approaches.
Low	Sampling methodology is only briefly discussed. Therefore, most sampling information is missing and likely to have a substantial impact on results. AND/OR The sampling methodology does not represent best sampling methods, protocols, or guidelines for the chemical and media of interest (*e.g.*, outdated (but still valid) sampling equipment or procedures, long storage durations). AND/OR There are some inconsistencies in the reporting of sampling information (*e.g.*, differences between text and tables in data source, differences between standard

Data Quality Rating	Metric Description
	method and actual procedures reported to have been used, etc.) which led to a low confidence in the sampling methodology used.
Critically Deficient	The sampling methodology is not discussed in the data source or companion source. AND/OR Sampling methodology is not scientifically sound or is not consistent with widely accepted methods/approaches for the chemical and media being analyzed (*e.g.*, inappropriate sampling equipment, improper storage conditions). AND/OR There are numerous inconsistencies in the reporting of sampling information, resulting in high uncertainty in the sampling methods used.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 2. Analytical methodology	
High	Samples were analyzed according to publicly available analytical methods that are scientifically sound and widely accepted (*i.e.*, from a source generally using sound methods and/or approaches) and are appropriate for the chemical and media of interest. Examples include EPA SW-846 Methods, NIOSH Manual of Analytical Methods 5th Edition, etc. OR The analytical method used was not a publicly available method from a source generally known to use sound methods and/or approaches, but the methodology is clear and appropriate (*i.e.*, scientifically sound) and similar to widely accepted protocols for the chemical and media of interest. All pertinent analytical ~~sampling~~ information is provided in the data source or companion source. Examples include: • extraction method • analytical instrumentation (required) • instrument calibration • LOQ, LOD, detection limits, and/or reporting limits • recovery samples • biomarker used (if applicable) • matrix-adjustment method (*i.e.*, creatinine, lipid, moisture)
Medium	Analytical methodology is discussed in detail and is clear and appropriate (*i.e.*, scientifically sound) for the chemical and media of interest; however, one or more pieces of analytical information is not described. The missing information is unlikely to have a substantial impact on results. AND/OR The analytical method may not be standard/widely accepted, but a method validation study was conducted prior to sample analysis and is expected to be consistent with sound scientific theory and/or accepted approaches. AND/OR

Data Quality Rating	Metric Description
	Samples were collected at a site and immediately analyzed using an on-site mobile laboratory, rather than shipped to a stationary laboratory.
Low	Analytical methodology is only briefly discussed. Analytical instrumentation is provided and consistent with accepted analytical instrumentation/methods. However, most analytical information is missing and likely to have a substantial impact on results. AND/OR Analytical method is not standard/widely accepted, and method validation is limited or not available. AND/OR Samples were analyzed using field screening techniques. AND/OR LOQ, LOD, detection limits, and/or reporting limits not reported. AND/OR There are some inconsistencies or possible errors in the reporting of analytical information (*e.g.*, differences between text and tables in data source, differences between standard method and actual procedures reported to have been used, etc.) which leads to a lower confidence in the method used.
Critically Deficient	Analytical methodology is not described, including analytical instrumentation (*i.e.*, HPLC, GC). AND/OR Analytical methodology is not scientifically appropriate for the chemical and media being analyzed (*e.g.*, method not sensitive enough, not specific to the chemical, out of date). AND/OR There are numerous inconsistencies in the reporting of analytical information, resulting in high uncertainty in the analytical methods used.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
<u>Metric 3.</u> Selection of biomarker of exposure	
High	Biomarker in a specified matrix is known to have an accurate and precise quantitative relationship with external exposure, internal dose, or target dose (*e.g.*, previous studies (or the current study) have indicated the biomarker of interest reflects external exposures). AND Biomarker (parent chemical or metabolite) is derived from exposure to the chemical of interest.
Medium	Biomarker in a specified matrix has accurate and precise quantitative relationship with external exposure, internal dose, or target dose. AND

Data Quality Rating	Metric Description
	Biomarker is derived from multiple parent chemicals, not only the chemical of interest, **but** there is a stated method to apportion the estimate to only the chemical of interest
Low	Biomarker in a specified matrix has accurate and precise quantitative relationship with external exposure, internal dose, or target dose. AND Biomarker is derived from multiple parent chemicals, not only the chemical of interest, **and** there is NOT a stated method to apportion the estimate to only the chemical of interest. OR Biomarker in a specified matrix is a poor surrogate (low accuracy and precision) for exposure/dose.
Critically Deficient	Not applicable. A study will not be deemed critically deficient based on the use of biomarker of exposure. ~~Biomarker in a specified matrix is a poor surrogate (low accuracy and precision) for exposure/dose.~~
Not Rated/Not Applicable	Metric is not applicable to the data source.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 2. Representative	
Metric 4. Testing scenario	
High	Testing conditions closely represent relevant exposure scenarios (*i.e.,* population/scenario/media of interest). Examples include: • amount and type of chemical/product used • source of exposure/test substance • method of application or by-stander exposure • use of exposure controls • microenvironment (location, time, climate, temperature, humidity, pressure, airflow) AND Testing conducted under a broad range of conditions for factors such as temperature, humidity, pressure, airflow, and chemical mass/weight fraction (if appropriate).
Medium	The data likely represent the relevant exposure scenario (*i.e.,* population/scenario/media of interest). One or more key pieces of information may not be described but the deficiencies are unlikely to have a substantial impact on the characterization of the exposure scenario. AND/OR If surrogate data, activities seem similar to the activities within scope.
Low	The data lack multiple key pieces of information, and the deficiencies are likely to have a substantial impact on the characterization of the exposure scenario. AND/OR

Data Quality Rating	Metric Description
	There are some inconsistencies or possible errors in the reporting of scenario information (*e.g.*, differences between text and tables in data source, differences between standard method and actual procedures reported to have been used, etc.) which leads to a lower confidence in the scenario assessed. AND/OR If surrogate data, activities have lesser similarity but are still potentially applicable to the activities within scope. AND/OR Testing conducted under a single set of conditions, except for experiments to determine a weight fraction or concentration in a product.
Critically Deficient	Testing conditions are not relevant to the exposure scenario of interest for the chemical.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 5. Sample size and variability	
High	Sample size is reported and large enough (*i.e.*, ≥ 10 samples) to be reasonably assured that the samples represent the scenario of interest. AND Replicate tests performed and variability across tests is characterized (if appropriate).
Medium	Sample size is moderate (*i.e.*, 5 to ~~10~~ <10 samples), thus the data are likely to represent the scenario of interest. AND Replicate tests performed and variability across tests is characterized (if appropriate).
Low	Sample size is small (*i.e.*, <5 samples), thus the data are likely to poorly represent the scenario of interest. AND/OR Replicate tests were not performed.
Critically Deficient	Sample size is not reported. AND/OR Single sample collected per data set, except for experiments to determine a weight fraction or concentration in a product. AND/OR For biomonitoring studies, the timing of sample collected is not appropriate based on chemical properties (*e.g.*, half-life), the pharmacokinetics of the chemical (*e.g.*, rate of uptake and elimination), and when the exposure event occurred.
Not Rated/Not Applicable	

Data Quality Rating	Metric Description
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 6. Temporality	
High	Source(s) of tested items appears to be current (within 5 years).
Medium	Source(s) of tested items is less consistent with when current or recent exposures (>5 to 15 years) are expected.
Low	Source(s) of tested items is not consistent with when current or recent exposures (>15 years) are expected or is not identified.
Critically Deficient	Temporality of tested items is not reported, discussed, or referenced.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 3. Accessibility/clarity	
Metric 7. Reporting of results	
High	Supplementary or raw data (*i.e.*, individual data points) are reported, allowing summary statistics to be calculated or reproduced. AND Summary statistics are detailed and complete. Example parameters include: • Description of data set summarized (*i.e.*, location, population, dates, etc.) • Range of concentrations or percentiles • Number of samples in data set • Frequency of detection • Measure of variation (CV, standard deviation) • Measure of central tendency (mean, geometric mean, median) • Test for outliers (if applicable) AND Both adjusted and unadjusted results are provided (*i.e.*, correction for void completeness in urine biomonitoring, whole-volume or lipid adjusted for blood biomonitoring) [only if applicable].
Medium	Supplementary or raw data (*i.e.*, individual data points) are not reported, and therefore summary statistics cannot be reproduced. AND/OR Summary statistics are reported but are missing one or more parameters (see description for high). AND/OR Only adjusted or unadjusted results are provided, but not both [only if applicable].

Data Quality Rating	Metric Description
Low	Supplementary data are not provided, and summary statistics are missing most parameters (see description for high). AND/OR There are some inconsistencies or errors in the results reported, resulting in low confidence in the results reported (*e.g.*, differences between text and tables in data source, less appropriate statistical methods).
Critically Deficient	There are numerous inconsistencies or errors in the calculation and/or reporting of results, resulting in highly uncertain reported results.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 8. Quality assurance	
High	The study applied quality assurance/quality control (QA/QC) measures and all pertinent QA/QC ~~quality assurance~~ information is provided in the data source or companion source. Examples include: • Laboratory, and/or storage recoveries. • Laboratory control samples. • Baseline (pre-exposure) samples. • Biomarker stability • Completeness of sample (*i.e.*, creatinine, specific gravity, osmolality for urine samples) AND No QA/QC ~~quality control~~ issues were identified, or any identified issues were minor and adequately addressed (*i.e.*, correction for low recoveries, correction for completeness).
Medium	The study applied and documented ~~quality assurance/quality control~~ QA/QC measures; however, one or more pieces of QA/QC information is not described. Missing information is unlikely to have a substantial impact on results. AND No QA/QC ~~quality control~~ issues were identified, or any identified issues were minor and addressed (*i.e.*, correction for low recoveries, correction for completeness).
Low	QA/QC ~~Quality assurance/quality control techniques~~ measures and results were not directly discussed but are ~~can be~~ implied through the study's use of standard field and laboratory protocols. AND/OR Deficiencies were noted in QA/QC ~~quality assurance/quality control~~ measures that are likely to have a substantial impact on results. AND/OR

Data Quality Rating	Metric Description
	There are some inconsistencies in the QA/QC ~~quality assurance~~ measures reported, resulting in low confidence in the ~~quality assurance/control~~ QA/QC measures taken and results (*e.g.*, differences between text and tables in data source).
Critically Deficient	QA/QC issues have been identified which significantly interfere with the overall reliability of the study.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 4. Variability and uncertainty	
Metric 9. Variability and uncertainty	
High	The study characterizes variability in the population/media studied. AND Key uncertainties, limitations, and data gaps have been identified. AND The uncertainties are minimal and have been characterized.
Medium	The study has limited characterization of variability in the population/media studied. AND/OR The study has limited discussion of key uncertainties, limitations, and data gaps. AND/OR Multiple uncertainties have been identified but are unlikely to have a substantial impact on results.
Low	The characterization of variability is absent. AND/OR Key uncertainties, limitations, and data gaps are not discussed. AND/OR Uncertainties identified may have a substantial impact on ~~the exposure~~ the exposure assessment
Critically Deficient	Estimates are highly uncertain based on characterization of variability and uncertainty.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

Table 5-3. Updated Data Evaluation Criteria for Database Data

Data Quality Rating	Description
Domain 1. Reliability	
Metric 1. Sampling methodology	
High	Widely accepted sampling methodologies (*i.e.,* from a source generally known to use ~~using~~ sound methods and/or approaches) were used to generate the data presented in the database. Example SOPs include USGS's "National Field Manual for the Collection of Water-Quality Data," EPA's "Ambient Air Sampling" (SESDPROC-303-R5), etc.
Medium	One or more pieces of sampling methodology information is not described, but missing information is unlikely to have a substantial impact on results. OR The sampling methodologies were consistent with sound scientific theory and/or accepted approaches based on the reported sampling information but may not have followed published procedures from a source generally known to use sound methods and/or approaches.
Low	The sampling methodology was not reported in data source or readily available companion data source.
Critically Deficient	The sampling methodologies used were not appropriate for the chemical/media of interest in the database (*e.g.,* inappropriate sampling equipment, improper storage conditions).
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 2. Analytical methodology	
High	Widely accepted analytical methodologies (*i.e.,* from a source generally using sound methods and/or approaches) were used to generate the data presented in the database. Example SOPs include EPA SW-846 Methods, NIOSH Manual of Analytical Methods 5th Edition, etc.
Medium	The analytical methodologies were consistent with sound scientific theory and/or accepted approaches based on the reported analytical information but may not have followed published procedures from a source generally known to use sound methods and/or approaches.
Low	The analytical methodology was not reported in data source or companion data source.
Critically Deficient	The analytical methodologies used were not appropriate for the chemical/media of interest in the database (*e.g.,* method not sensitive enough, not specific to the chemical, out of date).
Not Rated/Not Applicable	

Data Quality Rating	Description
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 2. Representative	
Metric 3. Geographic area	
High	Geographic location(s) is reported, discussed, or referenced.
Medium	Not applicable. This metric is dichotomous (*i.e.*, high vs. critically deficient).
Low	Not applicable. This metric is dichotomous (*i.e.*, high vs. critically deficient).
Critically Deficient	Geographic location is not reported, discussed, or referenced.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 4. Temporal	
High	The data reflect current conditions (within 5 years) AND/OR Database contains robust historical data for spatial and temporal analyses (if applicable).
Medium	The data are less consistent with current or recent exposures (>5 to 15 years) AND/OR Database contains sufficient historical data for spatial and temporal analyses (if applicable).
Low	Data are not consistent with when current exposures (>15 years old) may be expected AND/OR Database does not contain enough historical data for spatial and temporal analyses (if applicable).
Critically Deficient	Timing of sample data is not reported, discussed, or referenced.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 5. Exposure scenario	
High	The data closely represent relevant exposure scenario (*i.e.*, the population/scenario/media of interest). Examples include: • Amount and type of chemical/product used • Source of exposure • Method of application or by-stander exposure

Data Quality Rating	Description
	• Use of exposure controls • Microenvironment (location, time, climate)
Medium	The data likely represent the relevant exposure scenario (*i.e.*, population/scenario/media of interest). **One or more key pieces of information may not be described** but the deficiencies are unlikely to have a substantial impact on the characterization of the exposure scenario. AND/OR If surrogate data, activities seem similar to the activities within scope.
Low	The data lack multiple key pieces of information and the deficiencies are likely to have a substantial impact on the characterization of the exposure scenario. AND/OR There are **some inconsistencies or possible errors** in the reporting of scenario information (*e.g.*, differences between text and tables in data source, differences between standard method and actual procedures reported to have been used, etc.) which leads to a lower confidence in the scenario assessed. AND/OR If surrogate data, activities have lesser similarity but are still potentially applicable to the activities within scope.
Critically Deficient	If reported, the exposure scenario discussed in the monitored study does not represent the exposure scenario of interest for the chemical.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 3. Accessibility/clarity	
Metric 6. Availability of database and supporting documents	
High	Database is widely accepted and/or from a source generally known to use sound methods and/or approaches (*e.g.*, raw data from NHANES, STORET).
Medium	The database may not be widely known or accepted (*e.g.*, state-maintained databases), but the database is adequately documented with most or all of the following information: 1. Within the database, metadata is present (sample identifiers, annotations, flags, units, matrix descriptions, etc.) and-data fields are generally clear and defined. 2. A user manual and other supporting documentation is available, or there is sufficient documentation in the data source or companion source. Database quality assurance and data quality control measures are defined and/or a QA/QC protocol was followed.
Low	The database may not be widely known or accepted, and only limited database documentation is available (see the medium rating).

Data Quality Rating	Description
Critically Deficient	No information is provided on the database source or availability to the public.
Not Rated/ Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 7. Reporting of results	
High	The database or information source reporting the analysis of the database data is well organized and understandable by the target audience. AND Summary statistics in the data source are detailed and complete. Example parameters include: • Description of data set summarized (*i.e.*, location, population, dates, etc.) • Range of concentrations or percentiles • Number of samples in data set • Frequency of detection • Measure of variation (CV, standard deviation) • Measure of central tendency (mean, geometric mean, median) • Test for outliers (if applicable)
Medium	The database or information source reporting the analysis of the database data is well organized and understandable by the target audience. AND/OR Summary statistics are missing one or more parameters (see description for high).
Low	The database or information source reporting the analysis of the database data is unclear or not well organized. AND/OR Summary statistics are missing most parameters (see description for high) AND/OR There are some inconsistencies or errors in the results reported, resulting in low confidence in the results reported (*e.g.*, differences between text and tables in data source, less appropriate statistical methods).
Critically Deficient	There are numerous inconsistencies or errors in the calculation and/or reporting of results, resulting in highly uncertain reported results. AND/OR The information source reporting the analysis of the database data is missing key sections or lacks enough organization and clarity to locate and extract necessary information.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

Data Quality Rating	Description
Domain 4. Variability and uncertainty	
Metric 8. Variability and uncertainty	
High	Variability, key uncertainties, limitations, and/or data gaps have been identified. AND/OR The uncertainties are minimal and have been characterized.
Medium	The study has limited discussion of variability, key uncertainties, limitations, and/or data gaps. AND/OR Multiple uncertainties have been identified but are unlikely to have a substantial impact on results.
Low	Variability, key uncertainties, limitations, and data gaps are not discussed. AND/OR Uncertainties identified may have a substantial impact on the exposure the exposure assessment
Critically Deficient	Estimates are highly uncertain based on characterization of variability and uncertainty.
Not Rated/Not Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

5.5 Environmental and Human Health Hazard

Details regarding the evaluation and extraction of environmental and human health hazard information from references that met PECO screening criteria are available in Sections 5 and 6.4 of the 2021 Draft Systematic Review Protocol. Data quality criteria for environmental studies, animal and *in vitro* toxicity studies and epidemiological studies are available in Appendix P, Q, and R in the 2021 Draft Systematic Review Protocol, respectively (U.S. EPA, 2021). Any updates made to the data quality evaluation forms for human health hazard information since the 2021 Draft Systematic Review Protocol was published (U.S. EPA, 2021) are described below in Section 5.5.2. The below-listed supplemental documents provide details of the data evaluated and extracted. Data evaluation information for each discipline (*i.e.*, environmental and human health hazard) is contained in separate supplemental documents and includes metric ratings and the overall study quality determination for each data source. On the other hand, data extraction information for both disciplines are contained in a single supplemental document to increase the ease of accessing hazard data that may be relevant for both environmental- and human health-related receptors. One clarification that applies to the data extraction of human health hazard data is that all the data extraction was conducted in DistillerSR. In regard to the environmental hazard data, for references that meet PECO screening criteria at full text screening, the available environmental hazard data were extracted from those references in the ECOTOXicology Knowledgebase (ECOTOX) database and then imported into DistillerSR.

- *Data Quality Evaluation Information for Environmental Hazard for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025h)

- *Data Quality Evaluation Information for Human Health Hazard Epidemiology for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025k)
- *Data Quality Evaluation Information for Human Health Hazard Animal Toxicology for Butyl Benzyl Phthalate (BBP*) (U.S. EPA, 2025j)
- *Data Extraction Information for Environmental Hazard and Human Health Hazard Animal Toxicology and Epidemiology for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025b)

5.5.1 Environmental Hazard

As described in Appendix R of the 2021 Draft Systematic Review Protocol, references that met PECO criteria at full-text screening underwent data quality evaluation (U.S. EPA, 2021). Likewise, for references that met PECO criteria at full-text screening underwent data extraction as described in Section 6.4.1 of the Draft Systematic Review Protocol (U.S. EPA, 2021). This section describes any updates made to the data quality evaluation and data extraction process since the 2021 Draft Systematic Review Protocol was published.

For BBP, toxicity data gaps were identified for mammalian wildlife relevant to the terrestrial compartment of the environmental hazard assessment and thus rodent data for BBP were used as surrogate data for mammalian wildlife. The rodent data (TNO CIVO, 1993) were evaluated following the human health hazard animal toxicity evaluation process as described below in Section 5.5.2 and underwent data extraction as described in Section 6.4.1 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Additional data for health outcomes most relevant for environmental hazard assessment were also extracted for these rodent studies and are listed in detail in the *Data Extraction Information for Environmental Hazard and Human Health Hazard Animal Toxicology and Epidemiology for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025b).

Data Evaluation and Data Extraction Cross Walk

As per the established systematic review process described in the 2021 Draft Systematic Review Protocol, data extraction is completed for all health outcomes regardless of the overall quality determination a study has received during data quality evaluation (*i.e.*, rating of high, medium, low, or uninformative). Moreover, initial data extractions for environmental hazard are completed outside of DistillerSR by contractors that support ECOTOX, database managed by EPA's ORD. Data extraction QC for BBP was completed within DistillerSR by experts in environmental hazard data.

Since the 2021 Draft Systematic Review Protocol was published, an additional process improvement step has been incorporated into the environmental hazard TSCA systematic review process. Experts that perform the data extraction QC need to cross walk data evaluation forms to data extraction forms to ensure that health outcomes for each experimental condition reported in the study match in both the data evaluation and extraction forms; this step is necessary because the initial data extractions are completed outside of DistillerSR independently of the data evaluation process within DistillerSR. In addition, experts completing the cross walk during the data extraction QC need to ensure that the rating for the health outcome in the data evaluation forms is also reported in the data extraction forms.

To maximize efficiency for the completion of the data evaluation and data extraction cross walk, an external (outside of DistillerSR) automated function has been added. Figure 5-1 summarizes the steps that a study that meets the PECO screening criteria for environmental hazard (green circle in Figure 5-1) follows until completion of the data evaluation and data extraction cross walk (gray oval with check mark in Figure 5-1). The initial data extractions by ECOTOX contractors occur outside of DistillerSR (orange ECOTOX box in Figure 5-1), and data converted into a JSON file are later imported into

DistillerSR in preparation for the data extraction QC (second blue square in the red DistillerSR box in Figure 5-1).

The light purple box with the label "External processing" in Figure 5-1 illustrates the steps that occur outside of DistillerSR including the automated crosswalk function (blue square with an asterisk). Specifically, this automated function starts with a data extraction form and compares to the corresponding data evaluation form by first filtering by HERO ID, then filtering by species name, followed by lifestage of the organism, exposure duration, health outcome and chemical type. For each of these filtering levels as the matching function is run, if there is a data evaluation form that corresponds to the data extraction criteria, there is a successful match and the health outcomes in the data extraction form and data evaluation forms are aligned and, the rating is also added in the data extraction forms. On the contrary, if there is no data evaluation that corresponds to the data extraction criteria, the automated cross walk stops, and the outcome of the function is "No Match". If there is no match by the automated function, the cross walk is completed manually at the final step. Once the automated cross walk function is complete, the data are converted to a JSON file that is uploaded into DistillerSR. For the final step, the data extraction forms are reviewed for the successful automated matches and completes the cross walk manually for the forms that did not match (blue square with double asterisks in Figure 5-1), at which point the data evaluation and data extraction cross walk is complete.

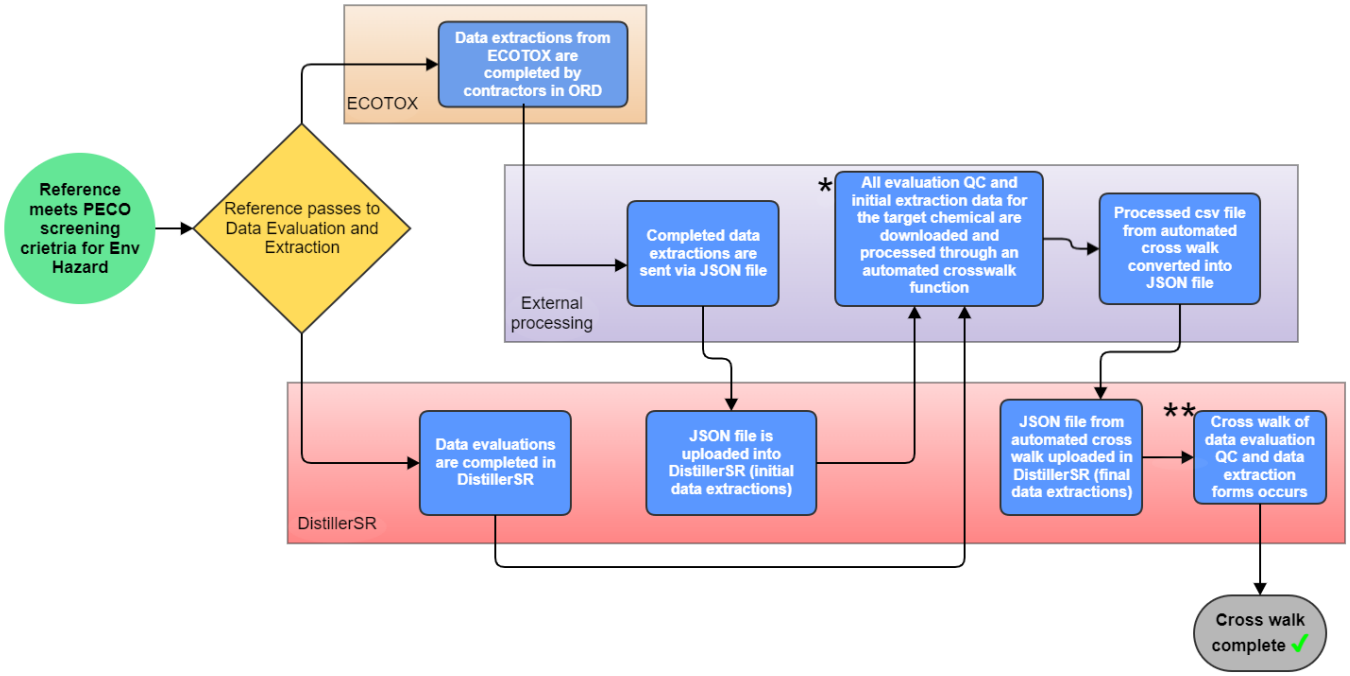


Figure 5-1. Data Evaluation and Data Extraction Cross Walk Workflow for Environmental Hazard

At the completion of the data evaluation and data extraction cross walk for BBP, the data extraction information was included in the *Data Extraction Information for Environmental Hazard and Human Health Hazard Animal Toxicology and Epidemiology for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025b).

5.5.2 Human Health Hazard

As described in Section 4.6.1, references that met further filtering criteria underwent data quality evaluation. This section describes updates made to the data quality evaluation and extraction forms since the 2021 Draft Systematic Review Protocol was published (U.S. EPA, 2021).

As a result of feedback from NASEM, the SACC, and multiple external stakeholders, OPPT explored ways to harmonize its Systematic Review Protocol with the IRIS Systematic Review Handbook. Besides being responsive to feedback, this effort was envisioned to have several additional benefits. It would facilitate the sharing of systematic review outputs between programs. This would not only make reviews reusable by other Agency units, but also could mean that chemical-specific assessments could be split up into modules, with each Agency unit sharing their results to form a final assessment. This in turn would conserve Agency resources. Harmonization of the protocols would also avoid waste of government funds (which is an imperative for all Agency managers) by not having employees and contractors in different EPA offices performing substantially similar reviews on the same references. Finally, it would prevent divergent conclusions from being reached by different parts of EPA within a very limited timeframe, supporting the vision of "One EPA".

The process of harmonizing the TSCA Systematic Review Protocol with the IRIS Systematic Review Handbook was a collaborative effort between OPPT and ORD. The OPPT team developed an IRIS/TSCA crosswalk that mapped corresponding IRIS and TSCA data quality evaluation domains. The IRIS data quality evaluation tool has fewer metrics compared to the old TSCA tool – an IRIS domain consisting of one metric might have a corresponding domain on the old TSCA form that consisted of several metrics; hence, multiple old TSCA metrics were mapped into a smaller number of IRIS metrics (many-to-one). Systematic review practitioners in both offices reviewed the mapping and confirmed that the data quality considerations on the old TSCA form were captured in the IRIS form. Therefore, new harmonized TSCA forms were developed based on the mapping of IRIS metrics to TSCA domains. Once general agreement was reached, a small number of references were used for calibration of the new forms to ensure 1) that the results were concordant between OPPT and IRIS and 2) that the results were concordant between the old TSCA data quality evaluation form and the harmonized data quality evaluation form. Once both the systematic review project managers and the teams of practitioner/evaluators were satisfied, the harmonized TSCA forms were finalized and put into use. Further details on the forms are described in the discipline-specific sections below (see Section 5.5.2.1 for details on the data evaluation forms for epidemiology studies and Section 5.5.2.2 for details on the data evaluation forms for animal toxicity studies used in assessing human health hazard).

5.5.2.1 Epidemiology Studies

As described above in Section 4.6.2.1, all references containing epidemiological information that met PECO screening criteria during full-text screening proceeded to an additional further filtering screening step. References that met the further filtering screening criteria then proceeded to data quality evaluation.

Epidemiology references that met the further filtering criteria were evaluated using the OPPT data quality evaluation form, which was modified to be more consistent with the IRIS data quality evaluation form, as described above. This modified OPPT form is referred to as the new harmonized TSCA epidemiology data quality evaluation form.

The old TSCA epidemiology data quality evaluation form used for other chemicals included 6 data quality evaluation domains, each of which included 3 or more metrics, such that the entire form included consideration of 22 different metrics. The new harmonized TSCA epidemiology data quality evaluation

form used for BBP includes the first 5 domains from the old TSCA data quality evaluation form, but the metrics are collapsed and streamlined with each domain having just one or two metrics. The new harmonized TSCA data quality evaluation form does not include the Biomarker domain from the old TSCA data quality evaluation form because biomarker considerations are now included in other domains. In particular, biomarkers of exposure are evaluated in Metric 2A of the Exposure Characterization Domain, biomarkers of effect are evaluated in Metric 3A of the Outcome Assessment Domain, and analytical components of biomarker assessments are evaluated in Metric 5A of the Analysis Domain. The evaluator assesses pre-defined criteria on the form to rate each metric as High, Medium, Low, or Critically Deficient for the reference.

The first step in developing the new harmonized data quality evaluation form was an IRIS-TSCA crosswalk that compared IRIS and TSCA domains, metrics, and criteria. Table 5-4 below summarizes the correspondence between IRIS and TSCA data quality evaluation domains. A more detailed crosswalk and discussion with experts from the ORD IRIS program indicated that all of the criteria that were assessed on the old TSCA form corresponded with components of the criteria assessed on the IRIS data quality evaluation form. Therefore, data quality evaluation criteria from the IRIS Handbook were used on the new harmonized TSCA forms. These criteria were further modified based on calibration discussions. The data quality evaluation instructions, domains, metrics, and criteria for the new harmonized TSCA Epidemiology Data Quality Evaluation form are presented below in Table 5-5.

The assessment of each of the metrics contributes to an OQD of High, Medium, Low, or Uninformative for the reference. Some references contain multiple health outcomes; therefore, a given reference may have multiple data quality evaluation forms and respective OQDs.

In addition to the updates to the data quality evaluation form, there were updates for data extraction. An update to the 2021 Draft Systematic Review Protocol is that the criteria for extracting data were refined. The criteria for extracting data from BBP epidemiology studies were that the reference met PECO screening criteria and further filtering criteria, and had an overall quality determination of High, Medium, or Low, and found statistically significant associations between BBP and an adverse health outcome. Additionally, the data extraction form for epidemiology studies was updated. Additional fields were added to the extraction form to facilitate evidence integration.

Table 5-4, Summary of Crosswalk of IRIS Domains, TSCA Domains, Old TSCA Form Metrics, and Harmonized TSCA Form Metrics for Epidemiology Studies

IRIS Domain (one metric per domain)	TSCA Domain	Old TSCA Form Metrics	Harmonized TSCA Form Domains and Metrics
Participant Selection	1. Study Participation	1, 2, 3	Domain 1, Metric 1A
Exposure Measurement	2. Exposure Characterization	4, 5, 6	Domain 2, Metric 2A
Outcome Ascertainment	3. Outcome Assessment	7. Outcome Measurement or Characterization	Domain 3, Metric 3A
Confounding	4. Potential Confounding / Variability Control	9, 10, 11	Domain 4, Metric 4A
Analysis	5. Analysis	12, 14, 15	Domain 5, Metric 5A
Selective Reporting	3. Outcome Assessment	8. Reporting Bias	Domain 3, Metric 3B
Sensitivity	5. Analysis	13. Statistical Power	Domain 5, Metric 5B
Part of other domains	6. Biomarkers	16, 17, 18, 19, 20, 21, 22	Part of Domains 2, 3, and 5

Table 5-5. Harmonized TSCA Epidemiology Data Quality Evaluation Form

Data Quality Rating	Description
Domain 1. Study participation (Combines/Collapses old TSCA Metrics 1, 2, and 3 into one metric - Metric 1A)	
Metric 1A. Participant Selection (Combines Old TSCA Form Metrics 1, 2, and 3)	
High	*Mark as high/good if:* *For all study types:* - There is minimal concern for selection bias based on description of recruitment process (*e.g.*, selection of comparison population, population-based random sample selection, recruitment from sampling frame including current and previous employees). - Exclusion and inclusion criteria for participants specified and would not induce bias. - Participation rate is reported at all steps of study (*e.g.*, initial enrollment, follow-up, selection into analysis sample). If rate is not high, there is appropriate rationale for why it is unlikely to be related to exposure (*e.g.*, comparison between participants and nonparticipants or other available information indicates differential selection is not likely).

Data Quality Rating	Description
Medium	*Mark as medium/adequate if:* - Enough of a description of the recruitment process to be comfortable that there is no serious risk of bias. - Inclusion and exclusion criteria for participants specified and would not induce bias. - Participation rate is incompletely reported but available information indicates participation is unlikely to be related to exposure.
Low	*Mark as low/deficient if:* - Little information on recruitment process, selection strategy, sampling framework and/or participation OR aspects of these processes raises the potential for bias (*e.g.*, healthy worker effect, survivor bias).
Critically deficient	*Mark as uninformative/critically deficient if:* - Aspects of the processes for recruitment, selection strategy, sampling framework, or participation result in concern that selection bias is likely to have had a large impact on effect estimates (*e.g.*, convenience sample with no information about recruitment and selection, cases and controls are recruited from different sources with different likelihood of exposure, recruitment materials stated outcome of interest and potential participants are aware of or are concerned about specific exposures).
Not rated/not applicable	*Mark as N/A if:* - Do not select for this metric.
Reviewer's comments	*Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance.*
Domain 2. Exposure characterization (Combines/Collapses old TSCA metrics 4, 5, and 6 into one metric – Metric 2A)	
Metric 2A. Exposure Measurement (Combines Old TSCA Form Metrics 4, 5, and 6)	
High	*Mark as high/good if:* - Valid exposure assessment methods were used, which represent the etiologically relevant time period of interest. - Exposure misclassification is expected to be minimal.
Medium	*Mark as medium/adequate if:* - Valid exposure assessment methods were used, which represent the etiologically relevant time period of interest. - Exposure misclassification may exist but is not expected to greatly change the effect estimate.
Low	*Mark as low/deficient if:* - Valid exposure assessment methods were used, which represent the etiologically relevant time period of interest. Specific knowledge about the exposure and outcome raise concerns about reverse causality, but there is uncertainty whether it is influencing the effect estimate. - Exposed groups are expected to contain a notable proportion of unexposed or minimally exposed individuals, the method did not capture important temporal or spatial variation, or there is other evidence of exposure misclassification that would be expected to notably change the effect estimate.

Data Quality Rating	Description
Critically deficient	*Mark as uninformative/critically deficient if:* - Exposure measurement does not characterize the etiologically relevant time period of exposure or is not valid. - There is evidence that reverse causality is very likely to account for the observed association. - Exposure measurement was not independent of outcome status. For Phthalates Only: For all short-chain phthalates and for primary metabolites (*e.g.*, MEHP, MINP) of long-chain phthalates and for phthalate diesters, if the only exposure measurement was a non-urine biomarker (*e.g.*, blood) then this metric should be rated as Uninformative/Critically Deficient. Biomarker matrices other than urine may be used for secondary metabolites of long-chain phthalates. (These criteria are based on the IRIS Protocol for the Systematic Review of the Health Effects of Phthalate Exposure, November 2017).
Not rated/not applicable	*Mark as N/A if:* - Do not select for this metric.
Reviewer's comments	*Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance.*
Domain 3. Outcome assessment (Includes corresponding IRIS metrics for old TSCA Metrics 7 and 8 – Metrics 3A and 3B, respectively)	
Metric 3A. Outcome Ascertainment (Corresponds to Old TSCA Form Metric 7. Outcome Measurement or Characterization)	
High	*Mark as high/good if:* - High certainty in the outcome definition (*i.e.*, specificity and sensitivity), minimal concerns with respect to misclassification. - Assessment instrument was validated in a population comparable to the one from which the study group was selected.
Medium	*Mark as medium/adequate if:* - Moderate confidence that outcome definition was specific and sensitive, some uncertainty with respect to misclassification but not expected to greatly change the effect estimate. - Assessment instrument was validated but not necessarily in a population comparable to the study group.
Low	*Mark as low/deficient if:* - Outcome definition was not specific or sensitive. - Uncertainty regarding validity of assessment instrument.
Critically deficient	*Mark as uninformative/critically deficient if:* - Invalid/insensitive marker of outcome. - Outcome ascertainment is very likely to be affected by knowledge of, or presence of, exposure. Note: Lack of blinding should not be automatically construed to be critically deficient.

Data Quality Rating	Description
Not rated/not applicable	*Mark as N/A if:* - Do not select for this metric.
Reviewer's comments	*Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance.*
Metric 3B. Selective Reporting (Corresponds to Old TSCA Form Metric 8. Reporting Bias) Note: It is currently rare that a study would cite a registered methods paper. Because we often can't know whether there is selective reporting, consistent with IRIS, this metric will often be rated as Medium/Adequate rather than Good/High. Ensure that the study's OQD is not getting downgraded from High to Medium solely because of the Selective Reporting Metric. But the metric itself will often be rated as Medium/Adequate.	
High	*Mark as high/good if:* - The results reported by study authors are consistent with the primary and secondary analyses described in a registered protocol or methods paper.
Medium	*Mark as medium/adequate if:* - The authors described their primary (and secondary) analyses in the methods section and results were reported for all primary analyses.
Low	*Mark as low/deficient if:* - Concerns were raised based on previous publications, a methods paper, or a registered protocol indicating that analyses were planned or conducted that were not reported, or that hypotheses originally considered to be secondary were represented as primary in the reviewed paper. - Only subgroup analyses were reported suggesting that results for the entire group were omitted. - Only statistically significant results were reported.
Critically deficient	*Mark as uninformative/critically deficient if:* - Do not select for this metric
Not rated/not applicable	*Mark as N/A if:* - Do not select for this metric.
Reviewer's comments	*Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance.*
Domain 4. Potential confounding/Variable control Potential Confounding / Variability Control (Combines/Collapses old TSCA metrics 9,10, and 11 into one metric – Metric 4A)	
Metric 4A. Potential Confounding (Combines Old TSCA Form metrics 9,10, and 11)	
High	*Mark as high/good if:* - Conveys strategy for identifying key confounders. This may include a priori biological considerations, published literature, causal diagrams, or statistical analyses; with recognition that not all "risk factors" are confounders. - Inclusion of potential confounders in statistical models not based solely on statistical significance criteria (*e.g.*, $p < 0.05$ from stepwise regression).

Data Quality Rating	Description
	- Does not include variables in the models that are likely to be influential colliders or intermediates on the causal pathway. - Key confounders are evaluated appropriately and considered to be unlikely sources of substantial confounding. This often will include: Presenting the distribution of potential confounders by levels of the exposure of interest and/or the outcomes of interest (with amount of missing data noted); Consideration that potential confounders were rare among the study population, or were expected to be poorly correlated with exposure of interest; Consideration of the most relevant functional forms of potential confounders; Examination of the potential impact of measurement error or missing data on confounder adjustment; Presenting a progression of model results with adjustments for different potential confounders, if warranted.
Medium	*Mark as medium/adequate if:* - Similar to high/good but may not have included all key confounders, or less detail may be available on the evaluation of confounders (*e.g.*, sub-bullets in high/good). It is possible that residual confounding could explain part of the observed effect, but concern is minimal.
Low	*Mark as low/deficient if:* - Does not include variables in the models that are likely to be influential colliders or intermediates on the causal pathway. And any of the following: - The potential for bias to explain some of the results is high based on an inability to rule out residual confounding, such as a lack of demonstration that key confounders of the exposure-outcome relationships were considered; - Descriptive information on key confounders (*e.g.*, their relationship relative to the outcomes and exposure levels) are not presented; or - Strategy of evaluating confounding is unclear or is not recommended (*e.g.*, only based on statistical significance criteria or stepwise regression [forward or backward elimination]).
Critically deficient	*Mark as uninformative/critically deficient if:* - Includes variables in the models that are colliders and/or intermediates in the causal pathway, indicating that substantial bias is likely from this adjustment; or - Confounding is likely present and not accounted for, indicating that all of the results were most likely due to bias.
Not rated/not applicable	*Mark as N/A if:* - Do not select for this metric
Reviewer's comments	*Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance.*
Domain 5. Analysis (Combines/Collapses old TSCA Metrics 12, 14, and 15 into one metric and includes the corresponding IRIS metric for TSCA Metric 13 – Metrics 5A and 5B, respectively)	

Data Quality Rating	Description
Metric 5A. Analysis (Combines Old TSCA Form Metrics 12, 14, and 15: Study Design and Methods, Reproducibility of Analyses, and Statistical Models)	
High	*Mark as high/good if:* - Use of an optimal characterization of the outcome variable. - Quantitative results presented (effect estimates and confidence limits or variability in estimates; *i.e.*, not presented only as a p-value or "significant"/ "not significant"). - Descriptive information about outcome and exposure provided (where applicable). - Amount of missing data noted and addressed appropriately (discussion of selection issues—missing at random vs. differential). - Where applicable, for exposure, includes LOD (and percentage below the LOD), and decision to use log transformation. - Includes analyses that address robustness of findings, *e.g.*, examination of exposure-response (explicit consideration of nonlinear possibilities, quadratic, spline, or threshold/ceiling effects included, when feasible); relevant sensitivity analyses; effect modification examined based only on a priori rationale with sufficient numbers. - No deficiencies in analysis evident. Discussion of some details may be absent (*e.g.*, examination of outliers).
Medium	*Mark as medium/adequate if:* Same as high/good except: - Descriptive information about exposure provided (where applicable) but may be incomplete; might not have discussed missing data, cut-points, or shape of distribution. - Includes analyses that address robustness of findings (examples in high/good), but some important analyses are not performed.
Low	*Mark as low/deficient if:* - Does not conduct analysis using optimal characterization of the outcome variable. - Descriptive information about exposure levels not provided (where applicable). - Effect estimate and p-value presented, without standard error or confidence interval. - Results presented as statistically "significant"/"not significant." - Sufficient details on test or model assumptions were not provided and there is some indication that the test or model might have been inappropriate.
Critically deficient	*Mark as uninformative/critically deficient if:* - Results of analyses of effect modification examined without clear a priori rationale and without providing main/principal effects (*e.g.*, presentation only of statistically significant interactions that were not hypothesis driven). - Analysis methods are not appropriate for design or data of the study.
Not rated/not applicable	*Mark as N/A if:* - Do not select for this metric.
Reviewer's comments	*Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance.*
Metric 5B. Sensitivity (Corresponds to Old TSCA Form Metric 13. Statistical Power)	
High	*Mark as high/good if:*

Data Quality Rating	Description
	- Study sensitivity was high due to sufficient exposure contrast, large sample size and examination of a relevant and sensitive population and minimal bias related to sensitivity in other domains.
Medium	*Mark as medium/adequate if:* - The range of exposure levels provides adequate variability to evaluate primary hypotheses in study. - The population was exposed to levels expected to have an impact on response. - The study population was sensitive to the development of the outcomes of interest (*e.g.*, ages, lifestage, sex). - The timing of outcome ascertainment was appropriate given expected latency for outcome development (*i.e.*, adequate follow-up interval). - The study was adequately powered to observe an effect, with a moderate sample size. - No other concerns raised regarding study sensitivity.
Low	*Mark as low/deficient if:* - Study sensitivity was deficient due to insufficient exposure contrast and/or small sample size in a non-sensitive or non-relevant population
Critically deficient	*Mark as uninformative/critically deficient if:* - There is a lack of critical information needed to inform the ability of the study to detect an effect if it exists, [and/or] there is indication that the study was unlikely to be able to do so.
Not rated/not applicable	*Mark as N/A if:* - Do not select for this metric.
Reviewer's comments	*Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance*
Overall Quality Determination (OQD)	
Additional Comments	Additional comments:
Based on your professional judgement, would you upgrade or downgrade this study's OQD?	*Select one of the following:* Yes, I would upgrade the paper Briefly describe why you decided to upgrade this study: Yes, I would downgrade the paper Briefly describe why you decided to downgrade this study: Neither – Keep quality rating as is
Specify which OQD you would give this paper (either confirm the auto calculated judgement OR suggest a new one based on your professional judgement?	High Medium Low Uninformative

5.5.2.2 Animal Toxicity Studies

Data quality evaluation of human health animal toxicity studies was conducted using the new harmonized data quality evaluation form. The impetus for development of this form was described above, the goal of which was to harmonize the data evaluation form from the existing TSCA Systematic Review Protocol with that from the IRIS Systematic Review Handbook. Table 5-6 describes the 6 domains and lists the number of metrics in each domain included in the new harmonized TSCA form. Since there are fewer domains in the IRIS Systematic Review Handbook than the TSCA Systematic Review Protocol, there was a many-to-one mapping from the old TSCA data quality evaluation form to the new harmonized TSCA data quality evaluation form as illustrated in the far-right column in Table 5-7. The far-right column depicts the individual metrics from the old TSCA data quality evaluation form that were mapped to the new harmonized TSCA data quality evaluation form. Moreover, Table 5-6 defines the domains in the new harmonized TSCA data quality evaluation form and describes how the old TSCA evaluation form metrics align with this new language. Detailed descriptions of each old TSCA form metrics in Table 5-6 can be found in Appendix Q of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021).

The new harmonized TSCA data quality evaluation form is described in Table 5-7 below. This form is applicable to the data quality evaluation of animal toxicity studies beyond BBP and thus will also be used in the systematic review of studies reporting exposure to other TSCA High Priority Substances.

With the impetus of preserving historic context and educate evaluators, explanatory text summarizing the origin of the new harmonized forms and how the old TSCA metrics map to the new harmonized TSCA domains in data evaluation forms can be found in the header row of Table 5-7. Extensive calibration sessions were completed to ensure the team of contractors and EPA staff were trained and confident that the two forms (*i.e.,* old TSCA form and harmonized TSCA form) produced equivalent results. Finally, all metrics in the data quality evaluation form include a comment box for reviewers to catalogue reference details not otherwise captured in the metric text, reading: “Reviewer comments: Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance.”

Table 5-6. Summary of Harmonized TSCA Domains and Domain Definitions, Harmonized TSCA Form Metrics, and Old TSCA Form Metrics for Human Health Animal Toxicity Studies.

Harmonized TSCA Form Domains	Harmonized TSCA Form Domain Definition	Harmonized TSCA Form Metrics	Old TSCA Form Metrics
Domain 1. Reporting quality	Domain 1 evaluates the reporting of details in the study. It uses two main categories of information: 1) critical, and 2) important. Critical information is considered essential and without it, the quality of the study may not be sufficiently evaluated. Important information is not required for evaluation, but it supports the critical information.	Single metric	Metrics 13, 14, and 15
Domain 2. Selection and performance	Domain 2 evaluates the risk of bias using metrics that assess allocation methods and observational bias. The randomization of the study design ensures that the effect observed is due to the exposure. Bias in observational measurements may lead to questions about the validity and reliability about the results of an experiment.	Metrics 2.1 and 2.2	Metrics 6 and 19
Domain 3. Confounding/Variable Control	Domain 3 evaluates the use of appropriate controls and/or comparators to discern the relationship between exposure to the test substance and the outcome(s)/endpoint(s) of interest. The use of controls and comparator and accounting for confounding variables minimizes bias so that the effect can be specifically attributed to the exposure.	Single metric	Metrics 4 and 5, 20, and 21
Domain 4. Selective Reporting and Attrition	Domain 4 evaluates the risk of bias due to selective reporting and attrition. The study should report intended sample sizes for all outcome(s)/endpoint(s) of interest, and discrepancies between the number of animals used to generate data points should also be adequately addressed. Attrition of animals during the experiment should be explained and transparent.	Single metric	Metric 22
Domain 5. Exposure methods sensitivity	Domain 5 evaluates the chemical administration and characterization. The information reported on the test substance should verify that exposure is in fact to the substance of interest, and the route and method of administration should be appropriate for the measured outcome(s)/endpoint(s) of interest. The timing, frequency, and duration of	Metrics 5.1 and 5.2	Metrics 1, 2, 3, 7, 8, 9, 10, and 12

Harmonized TSCA Form Domains	Harmonized TSCA Form Domain Definition	Harmonized TSCA Form Metrics	Old TSCA Form Metrics
	exposure should be suitable for all outcome(s)/endpoint(s) of interest.		
Domain 6. Outcome measures and results display	Domain 6 evaluates the sensitivity of the experiments that are used to characterize or measure the specific endpoint(s)/outcome(s) of interest. The methods used should reliably and reproducibly detect a response due to exposure for the specific endpoint(s)/outcome(s) of interest. The analysis and presentation of the results should be interpretable and transparent for the specific endpoint(s)/outcome(s) of interest.	Metrics 6.1 and 6.2	Metrics 11, 16, 17, 18, 23, and 24

Table 5-7. Harmonized TSCA Data Quality Evaluation Form for Human Health Animal Toxicity Studies

Data Quality Rating	Description
Domain 1. Reporting Quality (Combines Old TSCA Form Metrics 13, 14, and 15 from the Test Animals Domain)	
Does the study report information for evaluating the design and conduct of the study for the endpoint(s)/outcome(s) of interest? This Domain uses two main categories of information: 1) critical, and 2) important. Critical information necessary to perform study evaluation: Test animals' species, test article identity (*i.e.*, CASRN, chemical name, and/or structure), dose/concentration levels and duration of exposure, route (*e.g.*, oral; inhalation), qualitative or quantitative results for at least one endpoint of interest. Important information for evaluating the study methods: Test animal characteristics: source (*e.g.*, commercial source or laboratory-maintained colony), strain, age and/or life stage, sex, starting body weight, and/or parity (whether the test animals have been previously pregnant). For example, reporting animals to be 'mature' prior to starting the study leaves uncertainty and potential impact to results and may not be considered high quality. General animal husbandry conditions and procedures: temperature, humidity, light/dark cycle, diet, water availability, number of animals per cage throughout the study Exposure methods: test substance source, purity (or grade), method of administration Experimental design: frequency of exposure (*e.g.*, hours/day, days/week), number of animals per study group, animal age and life stage during exposure and at endpoint/outcome evaluation, as applicable to the study purpose/objective Endpoint evaluation methods: assays or procedures used to measure the endpoints/outcomes of interest. The presence or absence of all critical information determines whether a ranking is acceptable, or not. If/when critical information is missing, this Domain receives an uninformative ranking. The confidence level of acceptable, *e.g.*, high, medium, or low, corresponds to the amount of important information provided, in addition to the critical information. The confidence ranking for acceptable information should be justified and	

Data Quality Rating	Description
the assessor should identify which important information was provided in the study to support the assigned ranking. Note: This domain is limited to reporting. Other aspects (*i.e.*, appropriateness) of the exposure methods, experimental design, and endpoint evaluation methods are evaluated using the domains related to risk of bias and study sensitivity. The considerations below typically do not need to be refined by assessment teams, although in some instances the important information may be refined depending on the endpoints/outcomes of interest or the chemical under investigation. As for any study quality domain/metric, assessor judgment and rationale for ranking this domain should be given for the study and in the form of comments. Typically, a ranking given for this domain will not change across endpoints/outcomes investigated by the study. In the rationale, reviewers should indicate whether the study adhered to GLP, OECD, or other testing guidelines.	
High	*Mark as high/good if*: All critical and important information is reported or for the endpoints/outcomes of interest. The information could also be inferred from a reference document (*e.g.*, cited paper, manufacturer's website, guideline).
Medium	*Mark as medium/adequate if*: All critical information is reported but some combination important information is missing. However, the missing information is not expected to significantly impact the study evaluation.
Low	*Mark as low/deficient if*: All critical information is reported but important information is missing that is expected to significantly reduce the ability to evaluate the study.
Critically Deficient	*Mark as critically deficient if*: Study report is missing any pieces of critical information.
Not Rated/Not Applicable	*Mark as N/A if*: Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 2. Selection and Performance (Corresponds to Old TSCA Form Metrics 6 and 9)	
Metric 2.1. Allocation Were animals assigned to experimental groups using a method that minimizes selection bias? The considerations below typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each cohort or experiment in the study. Did each animal or litter have an equal/random chance of being assigned to any experimental group (*i.e.*, random allocation)? Is the allocation method described? Aside from randomization, were any steps taken to balance variables and/or pre-study test animal characteristics or other modifying factors across experimental groups during allocation? What is the expected and extent of the impact on study results if there is failure to randomize and/or normalize animal allocation? Is it significant or negligible?	

Data Quality Rating	Description
High	*Mark as high/good if*: Experimental groups were randomized, and any specific randomization procedure was described or inferable from a reference document (*e.g.*, cited paper, manufacturer's website, guideline). (*e.g.*, computer-generated scheme). Normalization of body weight to make sure average body weight is similar across doses if combined with a randomization scheme can be rated as *High*.
Medium	*Mark as medium/adequate if*: Authors report that groups were randomized but do not describe the specific procedure used (*e.g.*, "animals were randomized"). Alternatively, authors used a nonrandom method to control for important modifying factors across experimental groups (*e.g.*, body-weight normalization without use of randomization).
Low	*Mark as low/deficient if*: No indication of randomization of groups or other methods (*e.g.*, normalization) to control for important modifying factors across experimental groups.
Critically Deficient	*Mark as critically deficient if*: Bias in the animal allocations was explicitly reported or inferable from a reference document.
Not Rated/Not Applicable	*Mark as N/A if*: Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 2.2. Observational bias/Blinding Did the study implement measures to reduce observational bias? The considerations below typically do not need to be refined by the assessment teams. It is recommended that project assessors collectively build consensus to identify highly subjective measures of endpoints/outcomes where observational bias may strongly influence results prior to performing evaluations. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Does the study report blinding or other methods/procedures for reducing observational bias? This can apply to endpoints/outcomes that require heavy research practitioner handling or awareness of treatment/exposure groups during outcome assessment that may significantly impact study results. If not, did the study describe a design or approach for quality control of observational bias, for which such procedures can be inferred from a reference cited in the document? What is the expected and extent of the impact on study results of failure to implement (or report implementation) of these methods/procedures? Is it significant or negligible?	
High	*Mark as high/good if*: Measures to reduce observational bias were described (*e.g.*, blinding to conceal treatment groups during endpoint evaluation; consensus-based evaluations of histopathology-lesions).
Medium	*Mark as medium/adequate if*:

Data Quality Rating	Description
	Methods for reducing observational bias (*e.g.*, blinding) can be inferred from a cited reference (*e.g.*, cited paper or guideline) or were reported but were described incompletely. OR Measures to reduce observational bias were not described AND the potential concern for bias was mitigated because the outcomes were not subjective and/or based on use of automated/computer-driven systems, standard laboratory kits, simple objective measures (*e.g.*, body or tissue weight), or screening-level evaluations of histopathology.
Low	*Mark as low/deficient if*: Measures to reduce observational bias were not described AND the potential impact on the results is significant (*e.g.*, outcome measures are subjective).
Critically Deficient	*Mark as critically deficient if*: Strong evidence for observational bias that impacted the results.
Not Rated/Not Applicable	*Mark as N/A if*: Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
<u>Domain 3</u>. Confounding/Variable Control (Combines TSCA Metrics 4 and 5 from the Test Design Domain, Metric 20, and Metric 21 from the Confounding/Variable Control Domain)	
Are variables with the potential to confound or modify results controlled for and consistent across all experimental groups? The considerations below may need to be refined by assessment teams, as the specific variables of concern can vary by experiment or chemical. A judgment and rationale for this domain should be given for each cohort or experiment in the study, noting when the potential for confounding is restricted to specific endpoints/outcomes. Are there differences across the study groups (*e.g.*, co-exposures, vehicle, diet, palatability, husbandry) that could bias the results or introduce an unaccounted for or confounding variable? What is the expected extent of the impact on study results if confounding variables are identified? Is it significant or negligible?	
High	*Mark as high/good if*: Outside of the exposure of interest, variables that are likely to confound or modify results appear to be controlled for and consistent across experimental groups.
Medium	*Mark as medium/adequate if*: Some concern that variables that were likely to confound or modify results were uncontrolled or inconsistent across groups but are expected to have a minimal impact on the results.
Low	*Mark as low/deficient if*: Notable concern that potentially confounding variables were uncontrolled or inconsistent across groups and are expected to substantially impact the results.

Data Quality Rating	Description
Critically Deficient	*Mark as critically deficient if*: One or more confounding variables is known or presumed to be uncontrolled or inconsistent across groups and is expected to be a primary driver of the results and/or to distort the relationship between the exposure and outcome(s) of interest.
Not Rated/Not Applicable	*Mark as N/A if*: Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 4. Selective Reporting and Attrition (Combines TSCA Metric 22 from the Confounding/Variable Control Domain)	
Did the study report results for all prespecified outcomes and tested animals? Note: This domain does not consider the appropriateness of the analysis/results presentation. This aspect of study quality is evaluated in another domain. The considerations below typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each cohort or experiment in the study. Selective reporting bias: Are all results presented for endpoints/outcomes described in the methods? Attrition bias: Are all animals accounted for in the results? If there are discrepancies, do the authors provide an explanation (*e.g.*, death or unscheduled sacrifice during the study)? If unexplained results omissions and/or attrition are identified, what is the expected impact on the interpretation of the results?	
High	*Mark as high/good if*: Quantitative or qualitative results were reported for all prespecified outcomes (explicitly stated or inferred from a cited reference, such as a guideline or methodology peer-reviewed paper), exposure groups and evaluation time points. Data not reported in the primary article are available from supplemental material. If results omissions or animal attrition are identified, the authors provide an explanation, and these are not expected to impact the interpretation of the results.
Medium	*Mark as medium/adequate if*: Quantitative or qualitative results were reported for most prespecified outcomes (explicitly stated or inferred from a cited reference, such as a guideline or methodology peer-reviewed paper), exposure groups and evaluation time points. Omissions and/or attrition are not explained but are not expected to significantly impact the interpretation of the results.
Low	*Mark as low/deficient if*: Quantitative or qualitative results are missing for two or more prespecified endpoints (explicitly stated or inferred from a cited reference, such as a guideline or peer-reviewed methodology paper), exposure groups, and evaluation time points and/or there is high animal attrition; omissions and/or attrition are not explained and may significantly impact the interpretation of the results.

Data Quality Rating	Description
Critically Deficient	*Mark as critically deficient if*: Extensive results omission and/or animal attrition are identified and prevents comparisons of results across treatment groups.
Not Rated/Not Applicable	*Mark as N/A if*: Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 5. Exposure Methods Sensitivity (Combines TSCA Metrics from the Test Substance and Exposure Characterization Domains (Metrics 1,2,3,7,8,9,10, and 12))	
Metric 5.1. Chemical administration and characterization Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods? Was the route and method of exposure appropriate? Note: Relevance and utility of the routes of exposure are considered in the PECO criteria for study inclusion and during evidence synthesis. It is essential that the considerations below are considered, and potentially refined, by assessment teams, as the specific variables of concern can vary by chemical (*e.g.*, stability may be an issue for one chemical but not another). A judgment and rationale for this domain should be given for each cohort or experiment in the study. Are there concerns [specific to this chemical] regarding the source and purity and/or composition (*e.g.*, identity and percent distribution of different isomers) of the chemical? If so, can the purity and/or composition be obtained from the supplier (*e.g.*, as reported on the website)? Was independent analytical verification of the test article purity and composition performed? Did the authors take steps to ensure the reported exposure levels were accurate (*e.g.*, reporting by the authors of calculated doses in feeding/drinking water studies or sufficient information to independently calculate doses from concentrations in feed or water)? Are there concerns about the methods used to administer the chemical (*e.g.*, inhalation chamber type, gavage volume) or methods of test substance preparation or storage? For inhalation studies: Were target concentrations confirmed using reliable analytical measurements in chamber air? For oral studies: If necessary, based on consideration of chemical specific-knowledge (*e.g.*, instability in solution; volatility) and/or exposure design (*e.g.*, the frequency and duration of exposure), were chemical concentrations in the dosing solutions or diet/drinking water analytically confirmed? ** *If methods were cited to another publication, review the relevant methods in the original publication and consider this information as you rank this metric. Methods papers will be linked in HERO to the publication being evaluated.*	
High	*Mark as high/good if*: Chemical administration and characterization are complete (*i.e.*, test substance source and purity are appropriate, and analytic verification of the test article are provided). There are no concerns about the composition, stability, or purity of the administered chemical, or the specific methods of administration. For inhalation studies, chemical concentrations in the exposure chambers are verified using reliable analytical methods.

Data Quality Rating	Description
Medium	*Mark as medium/adequate if*: Some uncertainties in the chemical administration and characterization are identified but these are expected to have minimal impact on interpretation of the results (*e.g.*, source and vendor-reported purity are presented, but not independently verified; purity of the test article is suboptimal but not concerning; for inhalation studies with gases, actual exposure concentrations are missing or verified with less reliable methods; for oral and dermal studies, there are minor uncertainties about precision of dose levels or exposure concentrations).
Low	*Mark as low/deficient if*: Uncertainties in the exposure characterization are identified and are expected to substantially impact the results (*e.g.*, source of the test article was not reported; levels of impurities are substantial or concerning; deficient administration methods, such as use of static inhalation chambers or a gavage volume considered too large for the species and/or lifestage at exposure; for inhalation studies with aerosols or vapors, actual exposure concentrations are missing or verified with less reliable methods; for oral and dermal studies, there is substantial ambiguity about precision of dose levels or exposure concentrations).
Critically Deficient	*Mark as critically deficient if*: Uncertainties in the exposure characterization are identified and there is reasonable certainty that the results are largely attributable to factors other than exposure to the chemical of interest (*e.g.*, identified impurities are expected to be a primary driver of the results).
Not Rated/Not Applicable	*Mark as N/A if*: Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 5.2. Exposure timing, frequency, and duration Was the timing, frequency, and duration of exposure sensitive for the endpoint(s)/outcome(s) of interest? Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Does the exposure period include the critical window of sensitivity (*e.g.*, to detect developmental effects of interest)? Was the duration and frequency of exposure sensitive for detecting the endpoint of interest?	
High	*Mark as high/good if*: The timing, duration, and frequency of the exposure was sensitive, and the exposure included the critical window of sensitivity (if known).
Medium	*Mark as medium/adequate if*: The duration and frequency of the exposure was sensitive, and the exposure covered most of the critical window of sensitivity (if known).

Data Quality Rating	Description
Low	*Mark as low/deficient if*: The timing, duration, and frequency of the exposure is not sensitive or did not include most of the critical window of sensitivity (if known). These limitations are expected to bias the results towards the null.
Critically Deficient	*Mark as critically deficient if*: The exposure design is inappropriate for evaluating the outcome(s) of interest and is expected to strongly bias the results towards the null. The rationale should indicate the specific concern(s).
Not Rated/Not Applicable	*Mark as N/A if*: Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 6. Outcome Measures and Results Display (Combines TSCA Metrics from the Outcome Assessment and Data Presentation and Analysis Domains, and Metric 23 from the Data Presentation and Analysis Domain) (Metrics 11, 16, 17, 18, 23, and 24))	
Metric 6.1. Are the procedures sensitive and specific for evaluating the endpoint(s)/outcome(s) of interest? Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Are there concerns regarding the sensitivity, specificity, and/or validity of the protocols? Is the species appropriate? Are there serious concerns regarding the sample size? Are there concerns regarding the timing of the endpoint assessment? Examples of potential concerns include: Selection of protocols that are insensitive or nonspecific for the endpoint of interest Evaluations did not include all treatment groups (*e.g.*, only control and high dose) Use of unreliable methods to assess the outcome Assessment of endpoints at inappropriate or insensitive ages, or without addressing known endpoint variation (*e.g.*, due to circadian rhythms, estrous cyclicity) The study was conducted appropriately in relation to the evaluation domain, and any deficiencies, if present, are minor and would not be expected to influence the study results Decreased specificity or sensitivity of the response due to the timing of endpoint evaluation, as compared to exposure (*e.g.*, short acting depressant or irritant effects of chemicals; insensitivity due to prolonged period of non-exposure prior to testing) **** If methods were cited to another publication, review the relevant methods in the original publication and consider this information as you rank this metric. Methods papers will be linked in HERO to the publication being evaluated.*	
High	*Mark as high/good if*: The study was conducted appropriately in relation to the evaluation domain, and any deficiencies, if present, are minor and would not be expected to influence the study results.

Data Quality Rating	Description
Medium	*Mark as medium/adequate if*: There are methodological limitations relating to the evaluation domain, but that those limitations are not likely to be severe or have a notable impact on the results.
Low	*Mark as low/deficient if*: Biases or deficiencies were identified that are interpreted as likely to have had a notable impact on the results or that may prevent reliable interpretation of the study findings.
Critically Deficient	*Mark as critically deficient if*: The conduct of the study introduced a serious flaw that makes the observed effect(s) uninterpretable. Note: Sample size alone is not a reason to conclude an individual study is critically deficient.
Not Rated/Not Applicable	*Mark as N/A if*: Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 6.2. Results presentation Are the results presented in a way that makes the data usable and transparent? Considerations for this domain are highly variable depending on the outcomes of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Does the level of detail allow for an informed interpretation of the results? Are the data analyzed, compared, or presented in a way that is inappropriate or misleading? Examples of potential concerns include: Nonpreferred presentation (*e.g.*, developmental toxicity data averaged across pups in a treatment group, when litter responses are more appropriate; presentation of absolute organ-weight data when relative weights are more appropriate) Failing to present quantitative results either in tables or figures Lack of full presentation of the data (*e.g.*, presentation of mean without variance data; concurrent control data are not presented)	
High	*Mark as high/good if*: There was a full quantitative presentation of results (*e.g.*, means and SE or SD for continuous data; incidence data for categorical data; or individual animal results were presented). Any omissions are minor and are not expected to impact the interpretation of the results.
Medium	*Mark as medium/adequate if*: Some details of the results are missing, but the missing information is not expected to have a notable impact on the interpretation of the results.
Low	*Mark as low/deficient if*:

Data Quality Rating	Description
	Data were analyzed, compared, or presented in a way that is inappropriate or misleading (*e.g.*, the authors report a treatment-related effect on a quantitative endpoint, but only qualitative results are provided).
Critically Deficient	*Mark as critically deficient if*: Deficiencies in results presentation make the observed effect(s) uninterpretable.
Not Rated/Not Applicable	*Mark as N/A if*: Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Overall Quality Determination (OQD)	
Additional Comments	Additional Comments:
Based on your professional judgement, would you upgrade or downgrade this study's OQD?	*Select one of the following:* Yes, I would upgrade the paper Briefly describe why you decided to upgrade this study: Yes, I would downgrade the paper Briefly describe why you decided to downgrade this study: Neither – Keep quality rating as is
Specify which OQD you would give this paper (either confirm the auto calculated judgement OR suggest a new one based on your professional judgement?	High Medium Low Uninformative

5.6 Dermal Absorption

EPA's general approach to data evaluation and extraction of relevant data sources under TSCA is described in Sections 5 and 6, respectively of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). For each study, one reviewer conducts the initial review, and a second reviewer provides the QC review. EPA uses DistillerSR to evaluate and extract dermal absorption studies; the information from DistillerSR is then coded for output into tables that accompany the published risk evaluations. EPA evaluated and extracted dermal absorption studies that met the PECO screening criteria described above in Section 4.7.

Animal *in vivo* dermal absorption studies were evaluated using an extensively modified version of the animal toxicity data quality metrics shown in Appendix Q.4.2 of U.S. EPA (2021). To evaluate *in vitro/ex vivo* dermal absorption studies, EPA developed data evaluation metrics from the metrics used to evaluate *in vitro* mechanistic studies and presented a draft version of these metrics in Appendix S of the

2021 Draft Systematic Review Protocol (U.S. EPA, 2021). The sections below identify updates to these *in vivo* and *in vitro/ex vivo* criteria made since publication of the 2021 Draft Systematic Review Protocol.

Data extraction involves cataloguing experimental methods and results from the evaluated references. For *in vivo* studies, EPA extracts data on the matrices measured (*e.g.*, urine, carcass, exhaled air) and other information. For *in vitro* studies, EPA extracts information on the type of skin used (*e.g.*, source and area of body, thickness), the diffusion cell exposure set up (flow-through or static), and other data. For both *in vivo* and *in vitro/ex vivo* studies, EPA identifies the species used, whether skin was occluded, and information on the test substance and vehicle. As relevant, EPA extracts Kp/flux as well as fraction absorption information.

If adequate data are available from *in vivo* or *in vitro/ex vivo* (excised skin) studies, EPA will not evaluate, extract, or quantitively use data from the 3D human skin studies in risk evaluations. Currently, the 3D human skin equivalent models are not recommended by OECD Guidance (OECD Series on Testing and Assessment No. 156 (September 2022)) (OECD, 2022) for use in evaluating risks. However, EPA may discuss the 3D models when integrating evidence and may consider evaluating them if no other experimental dermal absorption information is available.

For BBP, EPA evaluated one *in vivo* rat and three *in vitro/ex vivo* studies (human and rat skin) from the literature searching and filtering of dermal absorption information. EPA assigned a medium OQD to the *in vivo* rat study. Rankings for *in vitro/ex vivo* studies were medium for two human skin studies and medium or uninformative for various experiments using human and rat skin in a third *in vitro/ex vivo* study. *Risk Evaluation for Butyl Benzyl Phthalate – Systematic Review Supplemental File: Data Quality Evaluation and Data Extraction Information for Dermal Absorption* (U.S. EPA, 2025d) provides details of the data extracted and evaluated, including metric rankings and the OQDs for evaluated data sources.

5.6.1 Data Quality Metrics- Animal *In Vivo*

Animal *in vivo* dermal absorption studies were evaluated using an extensively modified version of the animal toxicity data quality metrics shown in Appendix Q.4.2 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). The domains are identical except Domain 4 now refers to test models (instead of test *animals*). EPA used OECD guidelines to develop the criteria for the evaluation of *in vivo* dermal absorption references (OECD, 2011b, 2004a, b). Specifically, metrics were modified to address the standards used (metric 5), consistency of in exposure administration (metric 7), reporting of concentrations used (metric 8), exposure duration (metric 9), exposure groups and concentration (metric 10), characteristics of test animals and number of animal per group based on OECD 427 (metrics 11 and 13), outcome assessment methodology based on guidelines (metric 14), evaluation per group (metric 16), confounding variables (metric 17 and 18), data analysis, interpretation, and reporting (metric 19, 20, and 21). The full set of data quality metrics for *in vivo* animal studies are shown below.

Table 5-8. Data Quality Criteria for *In Vivo* Animal Dermal Absorption Studies

Data Quality Rating	Description
Domain 1. Test substance	
Metric 1. Test substance identity	

Data Quality Rating	Description
Was the test substance identified definitively (*i.e.*, established nomenclature, CASRN, physical nature, physical and chemical properties, and/or structure reported, including information on the specific form tested [*e.g.*, salt or base, valence state, isomer, if applicable] for materials that may vary in form)? If test substance was a mixture, were mixture components and ratios characterized?	
High	The test substance (*i.e.*, chemical of interest) was identified definitively (*i.e.*, nomenclature, CASRN, structure) and where applicable the specific form (*e.g.*, particle characteristics for solid state materials, salt or base, valence state, hydration state, isomer, radiolabel, etc.) was definitively and completely characterized. For mixtures, the components and ratios were characterized (*i.e.*, provided as concentration, ratio of percentage of the mixture or product). Additionally, for radiolabeled substances, the location of the radiolabel within the substance should be indicated, ideally with the radiolabel in a metabolically stable position
Medium	The test substance (*i.e.*, chemical of interest) was identified and the specific form was characterized (where applicable). For mixtures, some components and components and ratios were identified and characterized but at least the chemical of interest has a percentage/concentration reported. There were minor uncertainties (*e.g.*, minor characterization details were omitted such as about the radiolabel) that were unlikely to have a substantial impact on results
Low	The test substance and form (if applicable) were identified and the components and ratios of mixtures were characterized, but there were uncertainties regarding test substance identification or characterization that are likely to have a substantial impact on the results (*e.g.*, no information on isomer (or enantiomer) composition of differences could affect toxicokinetic properties, limited particle size information, omitted details regarding branched or straight chain structure).
Critically Deficient	The test substance identity and form (the latter if applicable) could not be determined from the information provided (*e.g.*, nomenclature was unclear and CASRN or structure were not reported) **OR** For mixtures, the components and ratios were not characterized.
Not Rated/Not Applicable	Do not select for this metric
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 2. Test substance source Was the source of the test substance reported, including manufacturer and batch/lot number for materials that may vary in composition? If synthesized or extracted, was test substance identity verified by analytical methods?	
High	The source of the test substance was reported as a manufacturer or the production process was specifically identified. The batch/lot number was identified (for materials that may vary in composition), and the chemical identity was either certified by the source in the publication or could be verified on a manufacturer's website. **OR** The test substance identity was analytically verified by the laboratory that performed the toxicity study.

Data Quality Rating	Description
Low	The test substance was synthesized or extracted by a source other than the manufacturer [and no production process was identified]. **OR** The source was not reported. **AND** The test substance identity was NOT analytically verified by the performing laboratory.
Not Rated/Not Applicable	Do not select for this metric
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 3. Test substance purity Was the purity or grade (*i.e.*, analytical, technical) of the test substance (including the radiolabeled substance) reported and adequate? Were impurities identified? Were impurities present in quantities that could influence the results? ~~Note that formaldehyde and other chemicals may require additional guidance that may differ from the guidance below.~~	
High	For discrete substances, the test substance purity **(including radiolabel)** and composition were such that any observed effects were highly likely to be due to the nominal test substance itself (*e.g.*, highly pure at >98% or analytical grade test substance or a formulation of lower purity that contains ingredients considered to be inert, such as water). ~~The radiopurity ideally should be greater than 95% and reasonable effort should be made to identify impurities present at or above 2%.~~ **AND** All components, including impurities and residual chemicals, were identified and the chemical of interest was the main component **(including the radiolabeled portion).**
Medium	The nature and quantity of reported impurities are such that study results were not likely to be substantially impacted by the impurities (impurities not known to induce outcome of interest at low levels, impurities are inert or GRAS, etc.). Regardless of the nature and purity, for discrete chemicals, the purity of the chemical of interest should be >70%, unless water is the only impurity.
Low	Purity and/or grade of test substance were not reported **(for both the labeled and unlabeled chemical).**
Critically Deficient	The nature and quantity of reported impurities (**for unlabeled and labeled substances)** were such that study results were likely to be due to one or more of the impurities. **AND/OR** For discrete chemicals, purity was <70% **(for unlabeled and labeled substances)** with an impurity other than water.
Not Rated/Not Applicable	Do not select for this metric
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

Data Quality Rating	Description
Domain 2. Test design	
Metric 4. Randomized allocation of animals Did the study explicitly report randomized allocation of animals to study groups?	
Medium	The study reported that animals were randomly allocated into study groups **OR** Allocation was performed with an unbiased method with a non-random component to ensure similar baseline characteristics across groups (*e.g.,* methods that account for body weight to ensure appropriate distribution across groups)
Low	The study did not report how animals were allocated to study groups, or there were deficiencies regarding the allocation method that are likely to have a substantial impact on results (*e.g.,* allocation by animal number).
Critically Deficient	The study reported using a biased method to allocate animals to study groups (*e.g.,* judgement of investigator). This is a serious flaw that makes the study unusable.
Not Rated/Not Applicable	Do not select for this metric
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 5. Standards for tests For assays with established criteria, were the test validity, acceptability, reliability, and/or QC criteria reported and consistent with current standards and guidelines? Were sufficient data provided to determine that the standards/guidelines have been met? **See Guidance for Reviewers to view examples of various criteria.** Example criteria: *Percent recovery:* 100±10% of the radioactivity as stated in OECD TG 427; 100±20% for volatile and unlabeled compounds as stated in OECD GD 28. *Coefficient of Variation*: **OECD 156 states that if the coefficient of variation is greater than 25%, then apply an adjustment.** ~~Variance across replicates should be measured and indicated when standard deviation exceeds 25%.~~	
Medium	Criteria used to determine the validity acceptability, reliability, and/or quality of the experiment (*e.g.,* percent recovery considered acceptable) were reported and consistent with current standards and guidelines, as/if applicable and authors stated that results met those criteria, or the results provided enough detail to compare with the criteria.
Low	Few or no QC criteria were reported, however, the reported results provided enough information to evaluate how the study compared against the criteria stated in the study and/or external criteria and standards.
Critically Deficient	Inadequate information was provided on the standards used to evaluate the study results **AND** 1) the authors did not report whether the test met pre-established criteria, **OR**

Data Quality Rating	Description
	2) inadequate data on results were presented to demonstrate the validity, acceptability, and reliability of the test when compared with current standards and guidelines or the pre-established standards/criteria identified by the authors. In this case, adequate QC cannot be performed.
Not Rated/Not Applicable	Do not select for this metric
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 3. Exposure characterization	
Metric 6. Preparation and storage of test substance (chemical) Did the study characterize preparation of the test substance and storage conditions? Were the frequency of preparation and/or storage conditions appropriate to the test substance stability and solubility (if applicable)?	
High	The test substance preparation and/or storage conditions (*e.g.*, test substance stability, homogeneity, mixing temperature, stock concentration, stirring methods, storage conditions) were reported and appropriate for the test substance and application scenario (*e.g.*, stability and solubility in diluents or solvents confirmed especially if they differ from what is used commercially; volatile test substances prepared and stored in sealed containers; same stock solution for all exposure concentrations).
Medium	The test substance preparation and storage conditions were reported, but minor limitations in the test substance preparation and/or storage conditions were identified (*e.g.*, test substance formulations were stirred instead of centrifuged for a specific number of rotations per minute). **OR** There is an omission of details that are unlikely to have a substantial impact on results (*e.g.*, preparation/administration of test substance is described, but storage of stock solution is not reported; however, storage is unlikely to affect results based on likely stability over the time frame of the test or the physical and chemical properties of the chemical make concerns about volatility or solubility unlikely).
	Deficiencies in reporting of test substance preparation, and/or storage conditions are likely to have a substantial impact on results (*e.g.*, available information on physical and chemical properties suggests that stability and/or solubility of test substance in diluent/solvent may be poor). **OR** Information on preparation and storage was *not* reported and lack of details could substantially impact results (*e.g.*, preparation for volatile or low-solubility chemicals).
Critically Deficient	Serious flaws reported regarding test substance preparation and/or storage conditions will have critical impacts on dose/concentration estimates and make the study unusable (*e.g.*, instability of test substance, test substance volatilized rapidly from storage containers).
Not Rated/Not Applicable	Do not select for this metric

Data Quality Rating	Description
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 7. Consistency of exposure administration Were exposures administered consistently across study groups (*e.g.*, consistent volumes/area of skin surface used for application that are ~ 5-10% of animal body surface (*e.g.*, 10 cm^2 for the rat), same area/location of body used for application)?	
High	Details of exposure administration were reported and exposures were administered consistently across study groups in a scientifically sound manner (*e.g.*, consistent volume and area of skin surface used for application, same area of body used for application for each animal and dose group).
Medium	Details of exposure administration were reported, but minor limitations in administration of exposures (*e.g.*, slight variations in surface area) were identified that are unlikely to have a substantial impact on results. **OR** Details of exposure administration are incompletely reported, but the missing information is unlikely to have a substantial impact on results.
Low	Details of exposure administration were reported, but deficiencies in administration of exposures (*e.g.*, moderate differences in of skin surface area used for application) that were reported or inferred from the text are likely to have a substantial impact on results. **OR** Details of exposure administration are insufficiently reported and the missing information is likely to have a substantial impact on results
Critically Deficient	Exposures were not administered consistently across and/or within study groups (*e.g.*, large differences in volume and area of skin surface used for application) resulting in serious flaws that make the study unusable.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 8. Reporting of concentrations Were exposure doses/concentrations or amounts of test substance applied to the skin reported without ambiguity (*e.g.*, point estimate instead of range, analytical instead of nominal, weight by weight vs volume by volume)? Note: Ambiguity also applies to doses/concentrations if values were only reported as points on a figure without numerical values.	
High	The exposure doses/concentrations or amounts of test substance were reported without ambiguity (*e.g.*, point estimate instead of range, analytical/measured instead of nominal, weight vs. volume).
Medium	The exposure doses/concentrations or amounts of test substance were reported with some ambiguity (*e.g.*, range instead of point estimate OR nominal instead of analytical/measured, unclear if weight or volume-based).

Data Quality Rating	Description
Low	The exposure doses/concentrations or amounts of test substance were reported but with substantial ambiguity about precision (*e.g.*, only an estimated range AND only nominal instead of analytical measurements).
Critically Deficient	The exposure doses/concentrations or amounts of test substance were not reported, resulting in serious flaws that make the study unusable.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 9. Exposure duration Was the exposure duration (*e.g.*, hours) reported and was it appropriate for this study type and/or outcome(s) of interest? Was the duration of exposure relevant to conditions of use and physical-chemical properties of the test substance? Did measurements continue post-exposure to account for retained dose in skin?	
High	The exposure duration (*e.g.*, hours) was reported and was appropriate based on the expected human exposure duration (typically at least 6 hours up to 24 hours following chemical application; if experiment continues beyond 1 day, measurements should continue daily in order to evaluate all excreta and tissues). A shorter exposure duration may also be included but is less useful unless the substance is volatile, the results demonstrate that absorption approached completion (*e.g.*, nothing left in the skin wash or tape strip samples), or the timepoint is used only for Kp/flux measurements.
Low	The duration(s) of exposure differed from current standards and guidelines for studies of this type (typically <6 to 24 hours prior to washing with excreta and/or measurements not continued without justification), and the differences may have a substantial impact on results.
Critically Deficient	No information on exposure duration(s) was reported **OR** the exposure duration was not appropriate **OR** Duration(s) differed significantly from studies of the same or similar types and these differences (most likely shorter duration) are likely to have a substantial impact on interpretation of results.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 10. Number of exposure groups and concentrations spacing Were the number of exposure groups/tested concentrations and dose/concentration spacing appropriate and justified by study authors (*e.g.*, to mimic a specific type of human exposure) and adequate for addressing the purpose of the study across a wide range of conditions of use (COUs) (*e.g.*, dilute, concentrated, and neat)?	
High	There were three or more dose groups tested and dose/concentration spacing were justified by study authors (*e.g.*, to mimic a specific type of human exposure) and were adequate for

Data Quality Rating	Description
	addressing the purpose of the study.
Medium	There were less than three group tested, however the choice of groups and diluent(s) were justified and are appropriate for common formulations. Any uncertainties given the reduced number of groups testes are minor relative to the difficulty of performing *in vivo* absorption testing.
Low	There were major limitations regarding the number of exposure groups and/or applied dose/concentration spacing (*e.g.,* dose and diluent testes are not very relevant to most exposure scenarios and only one dose/concentration tested), restricting the applicability of the results to only a subset of COUs and weight fractions.
Critically Deficient	The number of exposure groups and dose/concentrations spacing were not reported.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 4. Test model	
Metric 11. Test animal characteristics Were the animal species, strain, sex, age, and starting body weight reported? Was the test animal from a commercial source or in-house colony? Was the test species and strain an appropriate animal model for the evaluation of the specific(s) of interest (*e.g.,* routinely used for similar study types)? Per OECD 427, male rats of 200g -250g are suitable, particularly in the upper half of this range. The most sensitive sex should be used if there is evidence that one sex is more sensitive.	
High	The test animal species, strain, sex, age, and starting body weight were reported, and the test animal was obtained from a commercial source or laboratory-maintained colony. The test species and strain were an appropriate animal model for the evaluation of dermal absorption.
Medium	Minor uncertainties in the reporting of test animal characteristics (*e.g.,* age, or starting body weight) are unlikely to have a substantial impact on results. The test animals were obtained from a commercial source in-house colony, and the test species/strain/sex was an appropriate animal model for the evaluation of dermal absorption.
Low	The source or sex of the test animal was not reported. These deficiencies are likely to have a substantial impact on results. **OR** the test animal (species, strain, sex, life-stage, source) was not the best choice for the evaluation of dermal absorption.
Critically Deficient	The test animal species and any other necessary descriptive information were not at all reported.
Not Rated/Not	Do not select for this metric.

Data Quality Rating	Description
Applicable	
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 12. Adequacy and consistency of animal husbandry conditions	
High	All husbandry conditions were reported (*e.g.*, temperature, humidity, light-dark cycle, diet, water availability) and were adequate and the same for control and exposed populations, such that the only difference was exposure.
Medium	Most husbandry conditions were reported (see High bin) and were adequate and similar for all groups. Some differences in conditions were identified among groups, but these differences were considered minor uncertainties or limitations that are unlikely to have a substantial impact on results.
Low	Husbandry conditions were not sufficiently reported to evaluate if husbandry was adequate and whether differences occurred between control and exposed populations. These deficiencies are likely to have a substantial impact on results.
Critically Deficient	There were significant differences in husbandry conditions between control and exposed groups (*e.g.*, temperature, humidity, light-dark cycle). **OR** Animal husbandry conditions deviated from customary practices in ways likely to impact study results (*e.g.*, injuries and stress due to cage overcrowding). These are serious flaws that makes the study unusable.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 13. Number of animals per group Was the number of replicates per dose/concentration group appropriate for the study type and outcome analysis? OECD 427 states that "a group of at least four animals of one sex should be used for each test preparation and each scheduled termination time	
Medium	The number of animals per dose/concentration and timepoint group were reported and was appropriate (*e.g.*, acceptable data from a minimum of four animals per group, all from the same sex).
Low	The number of animals per dose/concentration and timepoint group was reported but was less than recommended by current standards and guidelines (*i.e.*, less than four animals tested or sexes were mixed). This is likely to have an impact on results. **OR** The number of replicates per dose/concentration was not reported.
Critically Deficient	The number of animals per study group was insufficient to characterize dermal absorption (*e.g.*, less than four replicates per test preparation produced acceptable data).

Data Quality Rating	Description
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 5. Outcome assessment	
Metric 14. Outcome assessment methodology Did the outcome assessment methodology address or report the intended absorption measurement of interest? Was the outcome assessment methodology (including measurement technique and timing of measurement[s]) appropriate for the associated conditions of use (COUs) and the dosing scenario? Were blood, urine, feces, and exhaled air (if necessary) individually collected at sampling time? [reference guidance notes re: infinite, nondepletable doses]	
High	The outcome assessment methodology addressed the intended absorption measurement AND was sensitive for the outcome(s) of interest and followed OECD guidance documents. The selected formulations are reasonable for the chemical of interest and would result in a sufficiently conservative estimate representative of conditions of use for the chemical of interest (*e.g.,* use of IPM as a diluent). All relevant bodily fluids were collected and measured. For percent absorption calculations finite dosing is required, normally 1-5 mg/cm^2 for a solid and up to 10 µL/cm^2 for liquids of test material, unless otherwise justified
Medium	The outcome assessment methodology used partially addressed the intended outcomes(s) of interest and deviations were explained, but minor uncertainties (*e.g.,* dosing was slightly below or above the recommendations for finite or infinite scenarios, did not assess all bodily fluids) are unlikely to have a substantial impact on results. If Kp determinations are presented, they should be from infinite dose or nondepletable conditions while finite dosing is required for percent absorption calculations. For infinite dose testing of solids, occlusion is required and at least 10 mg/cm^2 of pure substance must be used to establish an undepletable dose, regardless of concentration. For infinite dose testing of liquids/dilutions, occlusion is required, and flux must remain constant and steady-state throughout the duration of the experiment. Kp/flux measurements *in vivo* have substantial uncertainties; however a medium score can be achieved if efforts are taken to account for mass balance and ADME throughout the body (*e.g.,* shorter timepoints for measurement, collection of several tissues/excreta, see guidance notes).
Low	Significant deficiencies in the implementation of the reported outcome assessment methodology were identified (*e.g.,* a volatile diluent was used with a volatile test substance, etc.) **OR** The outcome assessment methodology was not clearly reported and it was unclear whether methods were sensitive for the outcome of interest. This is likely to have a substantial impact on results. For Kp/flux measurements, a low is assigned if efforts were not taken to account for potential missing absorbed dose through ADME processes (*e.g.,* only one tissue measured and/or

Data Quality Rating	Description
	delayed measurements that did not capture immediate absorption). Kp measurements are also downgraded if it is unclear whether the applied dose is non-depletable.
Critically Deficient	The reported assessment methodology was not sensitive to the outcome(s) of interest. For example, percentage absorption was determined only from an infinite dose, and/or Kp/flux was derived from a clearly finite dose, and statistics could not easily be calculated independently, or no relevant bodily fluids/tissues were assessed. These are serious flaws that make the study unusable.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 15. Consistency of outcome assessment Was the outcome assessment carried out consistently (*i.e.*, using the same protocol) across study groups (*e.g.*, assessment at the same time after initial exposure in all study groups)?	
High	Details of the outcome assessment protocol were reported, and outcomes were assessed consistently across study groups (*e.g.*, at the same time after initial exposure) using the same protocol in all study groups, the duration of exposure was the same across groups, the time periods when excreta were obtained were consistent across groups, etc.
Medium	There were minor differences in the timing of outcome assessment across study groups, or incomplete reporting of minor details of outcome assessment protocol execution were explained, but these uncertainties or limitations are unlikely to have substantial impact on results.
Low	Details regarding the execution of the study protocol for outcome assessment (*e.g.*, timing of assessment across groups) were confusing, limited, or not reported nor deviations explained, and these deficiencies are likely to have a substantial impact on results.
Critically Deficient	There were large inconsistencies in the execution of study protocols for outcome assessment across study groups. These are serious flaws that make the study unusable.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 16. Sampling adequacy and sensitivity Was the reported sampling size adequate for the outcome(s) of interest, including number of evaluations per exposure group, and endpoint (*e.g.*, scintillation counts/sample)?	
High	The study reported adequate sampling for the outcome(s) of interest including number of evaluations per exposure group, and measurement sensitivity (*e.g.*, scintillation counts/sample and/or duration of radioactivity detection, adequate signal to noise [*i.e.*, background] ratio for

Data Quality Rating	Description
	detection [*e.g.,* signal 3x noise]). The sampling intervals should be adequate to allow estimation of dermal absorption.
Medium	Details regarding sampling were reported, but minor limitations were identified in the reported sampling of the outcome(s) of interest and were explained. However, those limitations are unlikely to have a substantial impact on results.
Low	Details regarding sampling of outcomes were not fully reported nor explained and the omissions are likely to have a substantial impact on results.
Critically Deficient	Reported sampling was not adequate and/or serious uncertainties or limitations were identified in how the study carried out the sampling of the outcome(s) of interest (*e.g.,* replicates from control and test concentrations were evaluated at different times).
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 6. Confounding/variable control	
Metric 17. Confounding variables in test design and procedures Were there confounding differences among the study groups that could influence the outcome assessment (*e.g.,* differences in size of skin area exposed to the chemical, differences in test substance lot or batch that might have different purities)?	
High	There were no reported differences among study group parameters (*e.g.,* test substance lot or batch, initial starting weights) that could influence the outcome assessment.
Medium	Although the study did not report all information to determine whether confounding bias may exist, reported information did not identify differences (or identified only minor differences) among study groups in the above listed confounding factors. Minor differences were reported and explained in initial conditions that are unlikely to have a substantial impact on results.
Low	Reported information indicated moderate differences among the study groups with respect to body weight changes or other differences that may be attributed to systemic toxicity, or there were other major inconsistencies across study groups (*e.g.,* body weight variation was greater than 20% compared to mean).
Critically Deficient	There were significant differences among the study groups with respect to above considerations that make the data unreliable (*e.g.,* exposed skin was excessively hairy in one rodent compared to another, clear signs of damaged skin in some animals due to experimental procedures.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 18. Confounding variables in outcomes unrelated to exposure	

Data Quality Rating	Description
Were there differences among the study groups unrelated to exposure to test substance (*e.g.*, solubility in formulation) that could influence the outcome assessment? Were there differences among the study groups in animal attrition or health outcomes unrelated to exposure (*e.g.*, infection, damaged tissue) that could influence the outcome assessment? Professional judgement should be used to determine whether such differences would invalidate the study.	
High	There were no reported differences among the study animals or groups in test model unrelated to exposure (*e.g.*, solubility in formulation). Details regarding animal attrition and health outcomes unrelated to exposure (*e.g.*, infection, skin damage unrelated to treatment) were reported for each study group and there were no differences among groups that could influence the outcome assessment.
Medium	Authors reported that one or more animals or groups experienced disproportionate outcomes unrelated to exposure (*e.g.*, solubility issues, formulation-specific irritation), but data from the remaining exposure replicates or groups were valid and is unlikely to have a substantial impact on results. **OR** There was no information either to support or dismiss the suggestion that there were differences among groups in animal attrition, health outcomes unrelated to exposure, or solubility that could influence the outcome assessment.
Low	Data on outcome differences unrelated to exposure (*e.g.*, technical errors or variation in isolation of bodily fluids across test groups) were not reported for each study replicate or group and the missing information is likely to have a substantial impact on results.
Critically Deficient	There is evidence of insolubility in the formulation such that it was not properly demonstrating a diluted solution. **OR** Reported information indicated that study groups experienced attrition (*e.g.*, premature death) or health outcomes unrelated to exposure (*e.g.*, infection) that would render the full study (*i.e.*, all dose groups) unreliable considering the short-term duration.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 7. Data presentation and analysis	
Metric 19. Data analysis Were statistical methods, calculations methods, and/or data manipulation clearly described and appropriate for dataset(s)? Were absorption estimates **presented** ~~measured~~ across a time series for each compartment of the test system? **Did the results vary widely?**	
High	Statistical methods (including any calculations or data transformations) were clearly described or had only minor omissions and were appropriate for the dataset(s). Percent~~age~~ absorption estimates were measured across a time series for each compartment of the test system, and Kp/flux measurements were based on the linear/steady-state part of the absorption curve. Calculated absorption estimates properly accounted for outliers

Data Quality Rating	Description
	consistently across replicates/timepoints. **The coefficient of variation (CV) was ≤ 25% across samples, timepoints, dose groups in an individual experiment.**
Low	Statistical analysis was performed but not described adequately to understand what was performed or whether it was properly applied (*e.g.,* determination of outliers) **or** statistical analysis was inconsistently/inappropriately applied across replicates and datasets (*e.g.,* ~~absorption not measured across time series,~~ inconsistent exclusion of outliers (perhaps due to integrity failure) across measurements ~~but coefficient of variation for several replicates (SD relative to mean) was < 25%~~). **OR** Absorption estimates were not presented across a time series for each **scenario** ~~component~~. **OR** **[The CV was > 25% and ≤ 50% for more than half the samples across animals, replicates, media (e.g., receptor fluid, timepoints) within an individual scenario in a study.] OR [The CV was > 50% for more than half the samples within an individual scenario in a study, and data** are **available for EPA to calculate an alternate (upper end) value to account for variability in the results.]**
Critically Deficient	Statistical analysis was performed using an inappropriate method (*e.g.,* parametric test for non-normally distributed data) ~~and/or coefficient of variation for several replicates (SD relative to mean) was >25%.~~ **OR** Statistical analysis was not performed. **OR The coefficient of variation (CV) was >50% for more than half the samples (e.g., across samples, timepoints, dose groups) for an individual experiment.** **AND** Data enabling an independent statistical analysis **or to calculate an upper end value for fraction absorbed/Kp** were not provided. These are serious flaws that make the study unusable.
Not Rated/Not Applicable	Statistical analysis was not possible (n = 1−2) or not necessary **(clearly negative findings across all groups).**
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 20. Data interpretation Is the interpretation of results consistent with standards and guidelines? For example, did reported absorption estimates account for sufficient recovery? Was the combined amount of test substance in the skin (after removing appropriate tape strips if tape strips were used), blood, tissues, excreta, carcass and cage wash counted in the overall estimate? Was Kp vs fractional absorption results derived from the appropriate exposure conditions (infinite dose vs finite dose, respectively)?	
High	Recovery of applied test substance was adequate (mean of 100% +/- 10% or +/-20% for volatile chemicals; recoveries outside this range must be justified) or the absorption estimate was normalized to account for any reduction below these levels. Both the skin compartment and any tape-stripping washes after the first two were included in the absorption estimate.

Data Quality Rating	Description
	AND Assay results were correctly interpreted relative to the properties of the test substance and the assay setup (sufficient duration to capture all absorption if not evaporated, proper interpretation of finite vs infinite dose).
Medium	Absorption estimates were calculated improperly or incompletely (*e.g.*, skin compartment not included, values not normalized if recovery less than adequate), however simple independent data analysis is possible to overcome these issues.
Low	There are major uncertainties based on insufficient or incorrect interpretation of the results by the authors (*e.g.*, characterization of infinite vs finite doses), however EPA is able to estimate results with some level of confidence.
Critically Deficient	The reported scoring and/or evaluation criteria were very inconsistent with established practices, resulting in the interpretation of data results that are seriously flawed and highly misleading relative to the properly interpreted results (*e.g.*, study author claims 5% absorption but correct analysis results in 40% absorption; only percentage absorption **but not flux** is reported for **an infinite** ~~a finite~~ dose) and therefore not usable for any scenarios **AND** EPA is unable to confidently interpret the correct results based on the reported data.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 21. Reporting of data Were the data for all outcomes presented? Were data reported by exposure group? Per OECD 427, data should be presented as dislodgeable dose, skin compartment, blood concentration, excreta/expired air, and quantity remaining in carcass or removed organs. Irritation should also be reported if identified.	
High	Data for exposure-related findings were presented by exposure group (*e.g.*, all timepoints, formulations, concentrations, finite vs. infinite dose) and tissue compartments/bodily fluids of interest. Negative findings were reported qualitatively or quantitatively.
Medium	Data for exposure-related findings were reported for most, but not all, treatment levels (all tissue compartments/bodily fluids). The minor uncertainties in outcome reporting are unlikely to have substantial impact on results (*e.g.*, intermediate timepoints not included in the data tables but the full curve is included).
Low	Data for exposure-related findings were not shown for each treatment group, but results were described in the text. **OR** Data were reported inconsistently or with errors, however EPA was able to interpret the correct results with some level of confidence. **OR** Continuous data were presented without measures of variability or n/group.
Critically	Data presentation was inadequate (*e.g.*, the report does not differentiate among findings in

Data Quality Rating	Description
Deficient	multiple exposure groups) **OR** Major inconsistencies were present in reporting of results that render the findings unreliable and EPA is unable to confidently fill in gaps or make assumptions to make up for these uncertainties.
Not Rated/Not Applicable	Do not use for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

5.6.1 Data Quality Metrics - *In Vitro/Ex Vivo*

Table 5-9 presents the *in vitro/ex vivo* dermal absorption data evaluation criteria, as modified since publication of Appendix S of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). Language that was inserted is **bolded** and language removed is shown as ~~strikethrough~~. EPA used OECD guidelines to develop and update the criteria for the evaluation of *in vitro/ex vivo* dermal absorption references (OECD, 2011b, 2004a, c). For metrics 1, 3, 5, and 6 and 10-21, EPA made changes to the wording were made to provide context and/or clarity to the evaluation question and/or metric rankings. For metrics 4, 5, 7, 10 language was added in the places that were marked as TBD in Appendix S of U.S. EPA (2021). For metric 4, the wording originally used for the medium ranking was changed to indicate a high ranking and wording was added to the medium ranking. EPA also updated the low and critically deficient ranking descriptions. For metric 8, EPA removed the high ranking, and the description was incorporated into the medium ranking. EPA updated metric 19 to address data variability (the coefficient of variation) and revised metric 20 to clarify language and consider whether the reference calculated appropriate values (Kp/flux vs. fraction absorbed). The full set of *in vitro/ex vivo* data quality metrics are shown below.

Table 5-9. Updated Data Evaluation Criteria for *In Vitro*/*Ex Vivo* Dermal Absorption Studies

Data Quality Rating	Description
Domain 1. Test substance	
Metric 1. Test substance identity Was the test substance identified definitively (*i.e.*, established nomenclature, CASRN, physical nature, physical and chemical properties, and/or structure reported, including information on the specific form tested [*e.g.*, salt or base, valence state, isomer, if applicable] for materials that may vary in form)? If test substance was a mixture, were mixture components and ratios characterized?	
High	The test substance (*i.e.*, chemical of interest) was identified definitively (*i.e.*, nomenclature, CASRN, structure) and where applicable the specific form (*e.g.*, particle characteristics for solid state materials, salt or base, valence state, hydration state, isomer, radiolabel, etc.) was definitively and completely characterized. For mixtures, the components and ratios were characterized (*i.e.*, provided as concentration, ratio of percentage of the mixture or product). Additionally, for radiolabeled substances, the location of the radiolabel within the substance should be indicated, ideally with **the radiolabel** ^{14}~~C~~ in a metabolically stable position.
Medium	The test substance (*i.e.*, chemical of interest) was identified and the specific form was characterized (where applicable). For mixtures, some components and components and ratios were identified and characterized but at least the chemical of interest has a percentage/concentration reported. There were minor uncertainties (*e.g.*, minor characterization details were omitted **such as about the** radiolabel ~~details~~) that were unlikely to have a substantial impact on results.
Low	The test substance and form (if applicable) were identified, and the components and ratios of mixtures were characterized, but there were uncertainties regarding test substance identification or characterization that are likely to have a substantial impact on the results (*e.g.*, no information on isomer (or enantiomer) composition of differences could affect toxicokinetic properties, limited particle size information, omitted details regarding branched or straight chain structure).
Critically Deficient	The test substance identity and form (the latter if applicable) could not be determined from the information provided (*e.g.*, nomenclature was unclear and CASRN or structure were not reported) OR For mixtures, the components and ratios were not characterized.
Not Rated/Not Applicable	Do not select for this metric
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

Data Quality Rating	Description
Metric 2. Test substance source Was the source of the test substance reported, including manufacturer and batch/lot number for materials that may vary in composition? If synthesized or extracted, was test substance identity verified by analytical methods?	
High	The source of the test substance was reported as a manufacturer or the production process was specifically identified. The batch/lot number was identified (for materials that may vary in composition), and the chemical identity was either certified by the source in the publication or could be verified on a manufacturer's website. OR The test substance identity was analytically verified by the laboratory that performed the toxicity study.
Low	The test substance was synthesized or extracted by a source other than the manufacturer [and no production process was identified]. OR The source was not reported. AND The test substance identity was NOT analytically verified by the performing laboratory.
Not Rated/Not Applicable	Do not select for this metric
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 3. Test substance purity Was the purity or grade (*i.e.*, analytical, technical) of the test substance (including the radiolabeled substance) reported and adequate? Were impurities identified? Were impurities present in quantities that could influence the results?	
High	For discrete substances, the test substance **(including radiolabel)** purity and composition were such that any observed effects were highly likely to be due to the nominal test substance itself (*e.g.*, highly pure at >98% or analytical grade test substance or a formulation of lower purity that contains ingredients considered to be inert, such as water). All components, including impurities and residual chemicals, were identified and the chemical of interest was the main component **(including the radiolabeled portion)**.
Medium	The nature and quantity of reported impurities **(of the unlabeled and labeled portions of the chemical)** are such that study results were not likely to be substantially impacted by the impurities (impurities not known to induce outcome of interest at low levels, impurities are inert or GRAS, etc.). Regardless of the nature and purity, for discrete chemicals, the purity of the chemical of interest should be >70%, unless water is the only impurity.
Low	Purity and/or grade of test substance were not reported **(for both the labeled and unlabeled chemical).**

Data Quality Rating	Description
Critically Deficient	The nature and quantity of reported impurities **(for unlabeled and labeled substances)** were such that study results were likely to be due to one or more of the impurities. ~~This is a serious flaw that makes the study unusable.~~ **AND/OR** **For discrete chemicals, purity (for labeled and unlabeled substances) was <70% with an impurity other than water.**
Not Rated/Not Applicable	Do not select for this metric
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 2. Test design	
Metric 4. Reference compounds Were the results of a reference compound (*e.g.*, caffeine, testosterone, benzoic acid) run concurrently or separately and recently by the same laboratory and reported in the study? Was the absorption response appropriate? **Alternately, has the performing lab demonstrated previous technical sufficiency in dermal absorption studies?** ~~[TBD: need to decide how important it is to have reference compounds]~~	
High	**An appropriate concurrent reference compound was tested or data from a historical reference compound was provided, and an appropriate response was observed. Any uncertainties (*e.g.*, omission of minor details regarding exposure or response) are minor.**
Medium	~~When applicable, an appropriate concurrent or historical reference compound was used, and an appropriate response was observed. Any uncertainties (*e.g.*, omission of minor details regarding exposure or response) are minor.~~ **An appropriate concurrent or historical reference compound was used, but there were some deficiencies regarding the reference compound exposure or response (*e.g.*, the response was not well described, it is unclear whether the response was acceptable).**
Low	~~When applicable, an appropriate concurrent or historical reference compound was used, but there were deficiencies regarding the reference compound exposure or response (*e.g.*, the response was not described).~~ ~~OR~~ ~~No reference compound was used or reported.~~ **No appropriate reference compound was used or reported AND there is no established history of test performance in the performing laboratory.**
Critically Deficient	Reference compounds were run but ~~an inadequate response for the reference compounds (outside historical controls results~~) indicates ~~that the assay would not accurately measure absorption.~~ **the response was unacceptable (*e.g.*, outside historical control results), raising concerns about the validity of the assay.**
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

Data Quality Rating	Description
Metric 5. Assay procedures Were assay methods and procedures (*e.g.*, diffusion cell set up, temperature, humidity, physiological ~~conductivity~~ **compatibility** of receptor fluid, volumes applied ~~and surface area of skin~~, **amount of test substance per surface area of skin**, use/measurement of occlusion or carbon trap, materials and procedures used for tape stripping, capture of volatile compounds if required) described in detail and applicable/justified? ~~See other metrics for additional assay procedures (*e.g.*, metrics 1–3 for test substance information; metric 11 for exposure duration; metric 15 for replicates per group).~~ **Do the study methods describe how they ensure that quantification of the receptor fluid is adequately sensitive (*e.g.*, sufficient signal-to-noise ratio, high enough specific activity of radiolabel, sufficient amount of time or number of scintillations detected).** **Diffusion cell setup should indicate static vs flow-through, and for flow-through the flow rate should be indicated.** ~~OECD 428, OECD GD28 and OECD GD156 should be consulted and used to consider quality ratings.~~	
High	Study authors described the methods and procedures (*e.g.*, diffusion cell set up, temperature, humidity, physiological ~~conductivity~~ **compatibility** of receptor fluid, volumes applied and surface area of skin, use/measurement of occlusion or carbon trap, **specific activity of radiolabel**, materials and procedures used for tape stripping, capture of volatile compounds if required) used for the test in detail **and justified any relevant choices**. Either a static cell or flow-through system was used, with either constant stirring (static cell) or an appropriate flow- rate (flow-through). **These methods were appropriate based on the TGs and GDs above.**
Medium	Methods and procedures were partially described **(*e.g.*, all but temperature and humidity are described)** but appeared to be appropriate ~~(*e.g.*, TBD)~~, so the omission of details is unlikely to have a substantial impact on results.
Low	The methods and procedures were not well described or deviated from customary practices **(*e.g.*,** ~~TBD~~ **absence of occlusion or carbon trap for volatile test substance)** and this is likely to have a substantial impact on results, **however conservative statistical adjustments could possibly account for these deviations.**
Critically Deficient	Assay methods and procedures were not appropriate **and would result in unusable data that cannot be statistically accounted for** (*e.g.*, ~~TBD~~ **failure to use a diffusion cell with sufficient seal, too low volume/mass of test substance applied per surface area, tape stripping and wash fractions combined and not measured independently).**
Not Rated/Not Applicable	Do not select for this metric
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

Data Quality Rating	Description
Metric 6. Standards for tests For assays with established criteria, were the test validity, acceptability, reliability, and/or QC criteria reported and consistent with current standards and guidelines? **Were sufficient data provided to determine that the standards/guidelines have been met?** Example criteria: *Percent recovery:* 100±10% of the radioactivity as stated in OECD TG 428; 100±20% for volatile and unlabeled compounds as stated in OECD GD 28. ***Coefficient of Variation*: Variance across replicates should be measured and indicated when standard deviation exceeds 25%.** ***Skin integrity:* (1) Tritiated water – a.) a 'limit value' for a maximum Kp of 4.5 x10 $^{-3}$ cm/h (Guth et al. 2015 [Tox In Vitro 29:113-23]; Meidan and Roper, 2008 [Tox In Vitro 22:1062-9]) and mean Kp of 2.5 x 10-3 cm/h (Bronaugh et al. 1986 [Br J Dermatol 115:1-11]) for human *ex vivo* skin and b.) percent absorption (≤ 0.6% of applied dose in 1 hr) (Learn et al.– Poster from Charles River Labs).** **(2) Electrical conductance - minimal threshold of 17 kilo-ohms (Fasano et al., 2002) [Tox In Vitro 16:731-740]). (3) Trans-epidermal water loss - Less than 10 grams/m^2/hr (Zhang, 2018) [Tox In Vitro 51: 129-135] (4) Other internal reference standard methods (*e.g.*, 3H-labeled compounds, methylene blue) as cited in Guth et al. 2015.** **See Guidance for Reviewers to view examples of various criteria.** ~~*Skin integrity:* (1) Tritiated water – minimal flux threshold TBD (2) Electrical conductance - minimal threshold of 17 kilo-ohms (Fasano et al., 2002).~~ ~~OECD 428, OECD GD28, and OECD GD156 should be consulted; deviations should be explained.~~	
Medium	**Criteria used to determine the** ~~test~~ validity acceptability, reliability, and/or **quality of the experiment** ~~QC criteria~~ (*e.g.,* threshold for skin integrity, percent recovery considered acceptable) were reported and consistent with current standards and guidelines, as/if applicable **and authors stated that results met those criteria or the results provided enough detail to compare with the criteria**
Low	**Few or no QC criteria were reported, however, the reported results provided enough information to evaluate how the study compared against the criteria stated in the study and/or external criteria and standards.** ~~Some QC criteria were not reported~~.
Critically Deficient	Inadequate **information was provided on the standards used to evaluate the study results AND 1) the authors did not report whether the test met pre-established criteria,** **OR** 2**) inadequate** data **on results** were **presented** ~~provided~~ to demonstrate **the** validity, acceptability, and reliability of the test when compared with current standards and guidelines **or the pre-established standards/criteria identified by the authors. In this case, adequate QC cannot be performed.**
Not Rated/Not Applicable	Do not select for this metric
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

Data Quality Rating	Description
Domain 3. Exposure characterization	
Metric 7. Preparation and storage of test substance (chemical) Did the study characterize preparation of the test substance and storage conditions? Were the frequency of preparation and/or storage conditions appropriate to the test substance stability and solubility (if applicable)?	
High	The test substance preparation and/or storage conditions (*e.g.,* test substance stability, homogeneity, mixing temperature, stock concentration, stirring methods, storage conditions) were reported and appropriate for the test substance (*e.g.,* stability and solubility in diluents or solvents confirmed especially if they differ from what is used commercially; volatile test substances prepared and stored in sealed containers; same stock solution for all exposure concentrations).
Medium	The test substance preparation and storage conditions were reported, but minor limitations in the test substance preparation and/or storage conditions were identified (*e.g.,* **test substance formulations were stirred instead of centrifuged for a specific number of rotations per minute** ~~TBD~~). OR There is an omission of details that are unlikely to have a substantial impact on results (*e.g.,* preparation/administration of test substance is described, but storage is not reported; however, storage is unlikely to affect results based on likely stability over the time frame of the test or the physical and chemical properties of the chemical make concerns about volatility or solubility unlikely).
Low	Deficiencies in reporting of test substance preparation, and/or storage conditions are likely to have a substantial impact on results (*e.g.,* available information on physical and chemical properties suggests that stability and/or solubility of test substance in diluent/solvent may be poor). OR Information on preparation and storage was *not* reported and lack of details could substantially impact results (*e.g.,* preparation for volatile or low-solubility chemicals).
Critically Deficient	Serious flaws reported regarding test substance preparation and/or storage conditions will have critical impacts on dose/concentration estimates and make the study unusable (*e.g.,* instability of test substance, test substance volatilized rapidly from storage containers).
Not Rated/Not Applicable	Do not select for this metric
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 8. Consistency of exposure administration Were exposures administered consistently across study groups (*e.g.,* consistent volumes and area of skin surface for application)?	
~~High~~	~~Details of exposure administration were reported and exposures were administered consistently across study groups in a scientifically sound manner (*e.g.*, consistent volumes, thickness and area of skin surface for application,).~~

Data Quality Rating	Description
Medium	Details of exposure administration were reported or inferred from the text, **and** ~~but the minor limitations in administration of~~ exposures **were administered consistently across study groups in a scientifically sound manner** (*e.g.*, **consistent volumes** ~~slight variation in volume~~, thickness and area **of** ~~or~~ skin surface ~~used~~ for application). **Any minor deviations/limitations are considered)** ~~that were identified~~ are unlikely to have a substantial impact on results. OR Details of exposure administration are incompletely reported, but the missing information is unlikely to have a substantial impact on results.
Low	Details of exposure administration were reported, but deficiencies in administration of exposures (*e.g.*, moderate differences in volume, thickness, and area of skin surface used for application) that were reported or inferred from the text are likely to have a substantial impact on results. OR Details of exposure administration are insufficiently reported and the missing information is likely to have a substantial impact on results
Critically Deficient	Exposures were not administered consistently across and/or within study groups (*e.g.*, large differences in volume, thickness, and area of skin surface used for application) resulting in serious flaws that make the study unusable.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 9. Reporting of concentrations Were exposure doses/concentrations or amounts of test substance reported without ambiguity (*e.g.*, point estimate instead of range, analytical instead of nominal)? Note: Ambiguity also applies to doses/concentrations if values were only reported as points on a figure without numerical values.	
High	The exposure doses/concentrations or amounts of test substance were reported without ambiguity (*e.g.*, point estimate instead of range, analytical/measured instead of nominal).
Medium	The exposure doses/concentrations or amounts of test substance were reported with some ambiguity (*e.g.*, range instead of point estimate OR nominal instead of analytical/measured).
Low	The exposure doses/concentrations or amounts of test substance were reported but with substantial ambiguity about precision (*e.g.*, only an estimated range AND only nominal instead of analytical measurements).
Critically Deficient	The exposure doses/concentrations or amounts of test substance were not reported, resulting in serious flaws that make the study unusable.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

Data Quality Rating	Description
Metric 10. Exposure duration Was the exposure duration (*e.g.,* hours) reported and was it appropriate for this study type and/or outcome(s) of interest? **Was the duration of exposure relevant to conditions of use and physical-chemical properties of the test substance? Did measurements continue post-exposure to account for retained dose in skin?** ~~[TBD: add text about human exposure relevancy]~~.	
High	The exposure duration (*e.g.,* hours) was reported and was appropriate for the study type and/or outcome(s) of interest (*e.g.,* at least 6 to 10 hours prior to washing and up to at least 24 hours **total including** post-washing). A shorter exposure duration may also be included but is less useful unless the substance is demonstrated to be volatile, **the results demonstrate that absorption approached completion (*e.g.,* nothing left in the skin wash or tape strip samples),** or the timepoint is used only for Kp/flux measurements.
Low	The duration(s) of exposure differed slightly from current standards and guidelines for studies of this type (*e.g.,* <6 to 10 hours prior to washing and less than 24 hours **total including** post-washing), **and** ~~but~~ the differences **may** ~~are unlikely to~~ have a substantial impact on results.
Critically Deficient	No information on exposure duration(s) was reported OR the exposure duration was not appropriate OR Duration(s) differed significantly from studies of the same or similar types **and these differences (most likely shorter duration)** . ~~These deficiencies~~ are likely to have a substantial impact on **interpretation of** results.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 11. Number of exposure groups and concentrations spacing Were the number of exposure groups/**tested concentrations** and dose/concentration spacing **appropriate and** justified by study authors (*e.g.,* to mimic a specific type of human exposure) and adequate for addressing the purpose of the study **across a wide range of conditions of use (COUs) (*e.g.,* dilute, concentrated, and neat)?(***e.g.,* to evaluate dermal absorption)?	
High	**There were three or more dose** ~~The number of exposure~~ groups **tested** and dose/concentration spacing were justified by study authors (*e.g.,* to mimic a specific type of human exposure) and **were** ~~was~~ adequate for addressing the purpose of the study.
Low	There were ~~minor~~ limitations regarding the number of exposure groups and/or applied dose/concentration spacing (*e.g.,* unclear if lowest dose was low enough or the highest dose was high enough, **or less than three doses/concentrations tested), restricting the applicability of the results to only a subset of COUs and weight fractions**~~.), but the number of exposure groups and spacing of exposure levels were adequate and are unlikely to have a substantial impact on results.~~
Critically Deficient	The number of exposure groups and dose/concentration spacing were not reported ~~OR~~ ~~the number of exposure groups and dose/concentration spacing were not adequate and did not mimic expected human exposures.~~

Data Quality Rating	Description
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 4. Test model	
Metric 12. Test model **(skin)** Were the test models (*e.g.,* viable skin, cadaver/cosmetic surgery skin, **animal skin**) and descriptive information (*e.g.,* tissue origin, anatomical site, tissue storage, **initial** integrity or viability) reported? What was the source of the test model? Was the model routinely used for the outcome of interest? For example, for human skin, split thickness (200–400μm), dermatomed skin is preferred.	
High	The test model (*e.g.,* viable skin, cadaver skin, cosmetic surgery **skin, animal** skin) and descriptive information (*e.g.,* tissue origin, anatomical site, tissue storage, integrity or viability, **lot/batch used**) were reported and the test model was routinely used for the outcome of interest.
Low	The test model was **insufficiently** reported **and reporting** ~~along~~ with limited descriptive ~~information.~~ ~~OR~~ ~~The test model was routinely used for the outcome of interest. Reporting~~ limitations **may** ~~are unlikely to~~ have a substantial impact on results.
Critically Deficient	The test model and necessary descriptive information were not **at all** reported OR the test model was not appropriate for evaluation of the specific ~~outcome of interest~~
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 13. Number/Replicates per group Was the number of replicates per dose/concentration group appropriate for the study type and outcome analysis?	
Medium	The number of replicates per dose/concentration were reported and was appropriate (*e.g.,* acceptable data from a minimum of four replicates per test preparation).
Low	The number of replicates per dose/concentration **and timepoint** was reported but was less than recommended by current standards and guidelines **(*i.e.*, less than four replicates for each test preparation according to OECD TG 428)**. This is likely to have an impact on results. OR The number of replicates per dose/concentration was not reported.
Critically Deficient	The number of organisms or tissues per study group and/or replicates per study group was insufficient to characterize dermal absorption (*e.g.,* less than four replicates per test preparation produced acceptable data).

Data Quality Rating	Description
Not Rated/Not Applicable	**Do not select for this metric.** ~~Not applicable for qualitative studies not requiring any statistics.~~
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 5. Outcome assessment	
Metric 14. Outcome assessment methodology Did the outcome assessment methodology address or report the intended outcome(s) of interest? Was the outcome assessment methodology (including ~~nature of endpoints~~ ~~evaluated~~, measurement technique and timing of measurement[s]) **appropriate** ~~sensitive~~ for the **associated conditions of use (COUs)**~~outcome(s) of interest (*e.g.*, measured endpoints that are able to detect a true effect~~)? OECD 428, OECD GD28 and **the dosing scenario?** ~~OECD GD156 should be consulted, and deviations should be documented and explained.~~	
High	The outcome assessment methodology addressed the intended outcome(s) of interest AND was sensitive for the outcome(s) of interest and followed OECD guidance documents. The **selected formulations are reasonable for the chemical of interest and would result in** ~~dosing reflected~~ a **sufficiently conservative estimate representative** ~~range~~ of conditions of use **for the chemical of interest (*e.g.*, use of IPM diluent).** ~~(COUs) to which humans are exposed~~. The infinite dose scenario **should be used** ~~is optimum~~ for Kp determinations while finite dosing is **required** ~~optimal~~ for **percent**~~%~~ absorption calculations. **For finite** ~~The~~ dose **conditions, normally 1-5 mg/cm^2 of** ~~in the~~ **skin for a solid, and up to 10 µL/cm^2 for liquids of test material should be loaded, unless otherwise justified. For dilutions (*i.e.*, not neat test material), finite** ~~should be considered to be the potentially absorbable~~ dose **testing for each concentration of should ideally be conducted with application of 10 µL/cm^2 test material. For infinite dose testing of solids, it is required that at least 10 mg/cm^2 of pure substance be used to establish an undepletable dose, regardless of concentration. For infinite dose testing of liquids, at least 100 µL/cm^2 of pure substance should be used** to **establish an undepletable dose, regardless of concentration.** ~~calculate the final % absorption. Recovery is 90±10% or 80±20% for volatile substances.~~
Medium	The outcome assessment methodology used partially addressed the intended outcomes(s) of interest and deviations were explained, ~~(*e.g.*, mutation frequency evaluated in the absence of cytotoxicity in a gene mutation test)~~, but minor uncertainties **(*e.g.*, dosing was slightly below or above the recommendations for finite or infinite scenarios)** are unlikely to have a substantial impact on results.
Low	Significant deficiencies in the implementation of the reported outcome assessment methodology were identified (*e.g.*, **a volatile diluent was used with a volatile test substance** ~~matrix/assay interference, assay yielded anomalous results~~, etc.) **OR** The outcome assessment methodology was not clearly reported and it was unclear whether methods were sensitive for the outcome of interest. This is likely to have a substantial impact on results.
Critically Deficient	The reported assessment methodology was not sensitive to the outcome(s) of interest. For example, **percentage absorption was determined only from an infinite dose, and**/~~the reported measurement endpoint(s)~~ or **Kp/flux was derived from a finite dose, and statistics could** ~~timing~~

Data Quality Rating	Description
	~~were~~ not **easily be calculated independently.** ~~sensitive for the outcome(s) of interest (*e.g.*, cells were evaluated for chromosomal aberrations immediately after exposure to the test substance instead of after post-exposure incubation period).~~ These are serious flaws that make the study unusable.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 15. Consistency of outcome assessment Was the outcome assessment carried out consistently (*i.e.*, using the same protocol) across study groups (*e.g.*, assessment at the same time after initial exposure in all study groups)?	
High	Details of the outcome assessment protocol were reported and outcomes were assessed consistently across study groups (*e.g.*, at the same time after initial exposure) using the same protocol in all study groups. All study groups utilized the same **vehicle for the** blank formulation as **for the study concentration groups** ~~a vehicle~~, the duration of exposure was the same across groups, the same receptor fluid composition was **used** ~~utilized~~ for each group, the sampling period was consistent across groups, etc.
Medium	There were minor differences in the timing of outcome assessment across study groups, or incomplete reporting of minor details of outcome assessment protocol execution were explained, but these uncertainties or limitations are unlikely to have substantial impact on results.
Low	Details regarding the execution of the study protocol for outcome assessment (*e.g.*, timing of assessment across groups) were confusing, limited, or not reported nor deviations explained (or cited to another publication with no description in the paper itself), and these deficiencies are likely to have a substantial impact on results.
Critically Deficient	There were large inconsistencies in the execution of study protocols for outcome assessment across study groups. These are serious flaws that make the study unusable.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 16. Sampling adequacy **and sensitivity** Was the reported sampling size adequate for the outcome(s) of interest, including number of evaluations per exposure group, and endpoint (*e.g.*, **scintillation counts/sample)?**~~number of slides/cells/metaphases evaluated per test concentration)? OECD 428, OECD GD28, and OECD GD156 should be consulted, deviations should be explained.~~	
High	The study reported adequate sampling for the outcome(s) of interest including number of evaluations per exposure group, and **measurement sensitivity** ~~endpoint~~ (*e.g.*, scintillation counts/sample **and/or duration of radioactivity detection, adequate signal to noise [*i.e.*, background] ratio for detection [*e.g.*, signal 3x noise**]). The sampling intervals should be adequate to allow accurately graphically representing the results of the receptor fluid content of the test article versus time.

Data Quality Rating	Description
Medium	Details regarding sampling for the outcome(s) of interest were reported, but minor limitations were identified in the reported sampling of the outcome(s) of interest and were explained. However, those limitations are unlikely to have a substantial impact on results.
Low	Details regarding sampling of outcomes were not fully reported nor explained and the omissions are likely to have a substantial impact on results.
Critically Deficient	Reported sampling was not adequate for the outcome(s) of interest and/or serious uncertainties or limitations were identified in how the study carried out the sampling of the outcome(s) of interest (*e.g.*, replicates from control and test concentrations were evaluated at different times).
Not Rated/Not Applicable	**N/A** ~~NA~~ should be used for assays/studies that do not require a certain number of slides/cells/metaphases etc. be sampled for scoring (*i.e.*, mutagenicity assays, mechanistic studies).
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 6. Confounding/variable control	
Metric 17. Confounding variables in test design and procedures Were there confounding differences among the study groups in the size, and/or quality of tissues exposed that could influence the outcome assessment, **(*e.g.*, skin integrity)**?	
High	There were no differences reported among study group parameters (*e.g.*, test substance lot or batch, strain/batch/ lot number of ~~organisms~~ ~~or models used per group or size~~ **skin samples used per group or size**, and/or quality of tissues exposed) that could influence the outcome assessment. Skin integrity was ~~acceptable~~ **measured by preferable methods (*e.g.*, electrical resistance and TEWL). Results of skin integrity testing were acceptable for all replicates and exposure groups (*e.g.*, > 17 kilo-ohms based on electrical resistance, less than 10 grams/m2/hr)**
Medium	Minor differences were reported and explained in initial conditions that are unlikely to have a substantial impact on results (*e.g.*, tissues from two different lots were used and QC data were similar for both lots). Skin integrity ~~had variability but were acceptable~~ **was measured by a less desirable method (*e.g.*, tritiated water) , but results were acceptable (*e.g.*, a 'limit value' for Kp of 4.5 x10 $^{-3}$ cm/h or percent absorption of ≤ 0.6% of applied dose in 1 hr).** Outliers were statistically evaluated. **Most results of skin integrity testing were acceptable, and the number of replicates/donors was adequate after excluding any unacceptable results.**
Low	Initial strain/batch/lot number skin samples used per group, size, and/or quality of tissues exposed was not reported. These deficiencies are likely to have a substantial impact on results.
Critically Deficient	There were significant differences among the study groups with respect to the strain/batch/lot number of ~~organisms or models used per group or size and/or quality of tissues exposed (*e.g.*, initial number of viable bacterial cells were different for each replicate [105 cells in replicate 1, 108 cell in replicate 2, and 103 cells in replicate 3], tissues from two different lots were used for *in vitro* skin corrosion test, but the control batch quality for one lot was outside of the acceptability range). Skin integrity results were below thresholds. Recovery was below guidance limits or not quantified. Exposures did not reflect worker COUs.~~ **skin samples used per group or size and/or quality of tissues exposed (*e.g.*, several replicates demonstrated integrity issues)**.

Data Quality Rating	Description
	Recovery varied greatly among replicates (*i.e.*, >10%). In this situation, results are not reliable for estimating actual absorption.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 18. Confounding variables in outcomes unrelated to exposure Were there differences among the study groups unrelated to exposure to test substance (*e.g.*, **solubility in receptor fluid** ~~contamination~~) that could influence the outcome assessment? Did the test material interfere in the assay (*e.g.*, altering fluorescence or absorbance, signal quenching by heavy metals, altering pH, solubility, or stability issues)?	
High	There were no reported differences among the study replicates or groups in test model unrelated to exposure (*e.g.*, **solubility in receptor fluid** ~~contamination~~) and the test substance did not interfere with the assay (*e.g.*, signal quenching by heavy metals). The test substance was demonstrated to be soluble in the receptor fluid.
Medium	Authors reported that one or more replicates or groups experienced disproportionate outcomes unrelated to exposure (*e.g.*, **solubility issues** ~~contamination~~), but data from the remaining exposure replicates or groups were valid and is unlikely to have a substantial impact on results. OR The test material interfered in the assay, but the interference did not cause substantial differences among the groups. OR Solubility in the receptor fluid was not demonstrated, but solubility is not likely to be an issue based on the expected concentration relative to the receptor fluid formulation.
Low	Data on outcome differences unrelated to exposure (including receptor fluid formulation) were not reported for each study replicate or group and the missing information is likely to have a substantial impact on results. OR Assay interference was present or inferred resulting in large variabilities among the groups.
Critically Deficient	There were indications of assay interference several replicates or groups or there is evidence of insolubility in the receptor fluid such that no outcomes could be assessed.
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Domain 7. Data presentation and analysis	
Metric 19. Data analysis Were statistical methods, calculations methods, and/or data manipulation clearly described and appropriate for dataset(s)? **Were absorption estimates presented across a time series for each compartment of the test system? Did the results vary widely?**	

Data Quality Rating	Description
High	Statistical methods (including any calculations or data transformations) were clearly described or had only minor omissions and were appropriate for the dataset(s). Percent~~age~~ absorption estimates were presented across a time series for each compartment of the test system, and Kp/flux measurements were based on the linear/steady-state part of the absorption curve. **Calculated absorption estimates properly accounted for outliers consistently across replicates/timepoints. The coefficient of variation (CV) was ≤ 25% for more than half of the samples across each individual scenario (across donors, replicates, media (e.g., receptor fluid), timepoints) within the study.** ~~Any selection of outliers was justified.~~
Low	Statistical analysis was performed but not described adequately to understand what was performed or whether it was properly applied (*e.g.*, determination of outliers) **or** statistical analysis was inconsistently/inappropriately applied across replicates and datasets (*e.g.*, ~~absorption not measured across time series,~~ inconsistent exclusion of outliers **(perhaps due to integrity failure)** across measurements~~, coefficient of variation for several replicates (SD relative to mean) was <> 25%~~). **OR** **Absorption estimates were not presented across a time series for each scenario.** **OR** **[The CV was > 25% and ≤ 50% for more than half the samples across donors, replicates, media (e.g., receptor fluid, timepoints) within an individual scenario in a study**.] **OR [The CV was > 50% for more than half the samples within an individual scenario in a study, and data** are **available for EPA to calculate an alternate (upper end) value to account for variability in the results.]**
Critically Deficient	Statistical analysis was performed using an inappropriate method (*e.g.*, parametric test for non-normally distributed data)~~, and/or coefficient of variation for several replicates (SD relative to mean) was >25%.~~ OR Statistical analysis was not performed. **OR The coefficient of variation (CV) was >50% for more than half the samples (across donors, replicates, media (e.g., receptor fluid),** timepoints) **within an individual assay.** AND Data enabling an independent statistical analysis **or to calculate an upper end value for fraction absorbed/Kp** were not provided. These are serious flaws that make the study unusable.
Not Rated/Not Applicable	Statistical analysis was not possible (n = 1−2) or not necessary (clearly negative findings across all groups~~; Ames assay using 2-fold increase as benchmark~~).
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

Metric 20. Data interpretation

Is ~~Were the evaluation criteria reported and is~~ the interpretation of results consistent with standards and guidelines? **For example, did** reported absorption estimates account for sufficient recovery? Was the combined amount of test substance in the skin and receptor fluid counted in the overall estimate? **Was derivation of Kp vs fractional absorption applied to the appropriate exposure conditions (infinite dose vs finite dose, respectively)?**

Data Quality Rating	Description
High	~~Study authors followed evaluation criteria for the test, and these were consistent with established practices~~ [a]. Recovery of applied test substance was adequate (90% for occluded or non-volatile substance, 80% for non-occluded, volatile ~~substance~~ **or unlabeled substance) or the absorption estimate was normalized to account for any reduction below these levels. Both the skin compartment and any tape-stripping washes after the first two were included in the absorption estimate**. **AND** Assay results were correctly interpreted relative to the properties of the test substance and the assay setup (sufficient duration to capture all absorption if not evaporated, proper interpretation of finite vs infinite dose).
Medium	Absorption estimates were reported improperly **or incompletely** (*e.g.,* skin compartment not included, values not normalized if recovery less than adequate), however simple independent data analysis is possible to overcome these issues.
Low	**There are major uncertainties based on insufficient or incorrect interpretation of the results by the authors (e.g., characterization of infinite vs finite doses). However, EPA can estimate results with some level of confidence.** ~~Complex reanalysis of the data is required in order to obtain usable interpretations (*e.g.*, external outlier analysis may be required, Kp determination must be recalculated from the time series).~~
Critically Deficient	The reported **scoring** ~~rating~~ and/or evaluation criteria were very inconsistent with established practices, resulting in the interpretation of data results that are seriously flawed and highly misleading relative to the properly interpreted results (*e.g.,* study author claims 5% absorption but correct analysis results in 40% absorption, only percentage absorption is reported from a finite dose) **and therefore not usable for any scenarios.**
Not Rated/Not Applicable	Do not select for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*
Metric 21. Reporting of data Were the data for all outcomes presented? Were data reported by exposure group?	
High	Data for exposure-related findings were presented for all outcomes by exposure group ***(e.g., all timepoints, formulations, concentrations, finite vs infinite dose)***. Negative findings were reported qualitatively or quantitatively.
Medium	Data for exposure-related findings were reported for most, but not all, outcomes by exposure group (*e.g.,* both short and long-term exposures). The minor uncertainties in outcome reporting are unlikely to have substantial impact on results ***(e.g., intermediate timepoints not included in the data tables but the full curve is included).***
Low	Data for exposure-related findings were not shown for each study group, but results were described in the text. OR Data were only reported for some outcomes. OR Continuous data were presented without measures of variability or n/group.

Data Quality Rating	Description
Critically Deficient	Data presentation was inadequate (*e.g.*, the report does not differentiate among findings in multiple exposure groups) **OR** Major inconsistencies were present in reporting of results that render the findings uncertain regarding hazard identification or dose- response.
Not Rated/Not Applicable	Do not use for this metric.
Reviewer's Comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*

6 EVIDENCE INTEGRATION

As described in Section 7 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021), evidence integration refers to the consideration of evidence obtained from systematic review and scientific information obtained from sources that did not undergo systematic review to implement a weight of the scientific evidence approach. The weight of the scientific evidence is defined as "a systematic review method, applied in a manner suited to the nature of the evidence or decision, that uses a pre-established protocol to comprehensively, objectively, transparently, and consistently identify and evaluate each stream of evidence, including strengths, limitations, and relevance of each study and to integrate evidence as necessary and appropriate based upon strengths, limitations, and relevance" (40 CFR 702.33). The consideration of the quality and relevance of the data, while taking into account the strengths and limitations of the data, to appropriately evaluate the evidence for this supplement, is described in Section 7 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021).

6.1 Physical and Chemical Properties

The systematic review process identified multiple data for each of the physical-chemical properties analyzed in the risk evaluation. Relevant data types used for the physical-chemical assessment are discussed in Appendix K of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). When a specific datum is cited for a given physical-chemical parameter, priority is given to data from expert-curated, peer-reviewed databases that have been identified as "trusted sources". Sources of uncertainty are discussed, when appropriate, in the risk evaluation.

6.2 Environmental Fate and Transport

Relevant data types used for environmental fate and transport assessment are listed in Table 7-1 of the Draft Systematic Review Protocol (U.S. EPA, 2021). Systematic review data as well as data gaps filled using evidence streams outside systematic review are incorporated as described in Figure 7-1. Quality of these data are determined based on whether they are measured or estimated data, and further broken down based on consistency, study design, study conditions and uncertainty (Figure 7-2).

6.3 Environmental Release and Occupational Exposure

To evaluate environmental releases and occupational exposures for the various COUs, EPA first mapped the COUs to broader occupational exposure scenario (OES) categories, as shown in the *Environmental Release and Occupational Exposure Assessment for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025n). Specifically, EPA developed OES categories to group processes or applications with similar sources of environmental releases and occupational exposures. For each OES, EPA integrated the occupational exposure results for various job classifications to be representative of all U.S. workers and sites within that OES.

The EPA utilized programmatic release data from DMR. The EPA did not utilize release data from any other programmatic databases (such as the TRI and NEI databases), because BBP release reporting was not required and no data for BBP were reported. As a result, EPA used data from the systematic review literature, Emission Scenario Documents (ESDs), Generic Scenarios (GSs), and Specific Environmental Release Categories (SpERCs) to determine model input parameters for each OES. As described in the *Release and Occupational Exposure Assessment for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025n)., EPA ran Monte Carlo simulations with 100,000 iterations and the Latin Hypercube sampling method, using the statistical distribution for each input parameter to calculate a full distribution of the final release results for each OES. EPA selected the 50th and 95th percentiles of the resulting distributions to represent central tendency and high-end releases, respectively. To estimate the number of sites using BBP within an OES, EPA relied on the Chemical Data Reporting (CDR) (U.S. EPA, 2020a) database for

manufacturing and import sites. For all other OESs, EPA used GS and ESD inputs to estimate the number of sites and used U.S. Census Bureau data where necessary to provide a bounding estimate.

EPA assessed OES-specific exposures to workers and occupational non-users (ONUs) based on monitoring data, surrogate monitoring data, and modeling approaches. EPA developed worker activity information using GSs, ESD, SpERCs and other systematic review literature, as described in the *Environmental Release and Occupational Exposure Assessment for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025n). When sufficient monitoring data for an OES were available, EPA gave preference to monitoring data under 20 years old. Dermal exposure data were not available for any of the OES considered in this assessment. As a result, EPA modeled dermal loading using a flux-limited absorption model, which is further discussed in Section 6.6 of this document.

For OES where monitoring data or surrogate data were not available, EPA utilized literature and relevant ESDs, GSs, and SpERCs to determine input parameters and approaches to model the defining exposure activity for each OES. The application of adhesives and sealants and the application of paints and coatings OESs utilized the *Automotive Refinishing Spray Coating Mist Inhalation Model*. This model incorporates EPA-collected, surrogate spray application data obtained through a search of available OSHA *In-Depth Surveys of the Automotive Refinishing Shop Industry* and other relevant studies (OECD, 2011a). The *Environmental Release and Occupational Exposure Assessment for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025n) describes all models, approaches, and parameters. Where available, EPA used literature data to estimate the number of exposure days. EPA relied on U.S. Census Bureau data and OES-assigned NAICS codes to estimate the number of workers and ONUs potentially exposed to BBP within each OES.

6.4 General Population, Consumer, and Environmental Exposure

Butyl benzyl phthalate (BBP) concentrations in ambient air, surface water, sediment, soil, landfills, and biosolids were gathered and summarized within each environmental media pathway within the *Environmental Media and General Population and Environment Exposure for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025m). The sources and approaches to gather monitoring data from peer-reviewed publications, government reports, and/or databases were classified as monitoring and mainly used to compare with modeling results or to support qualitative assessments. Consumer products containing BBP were identified through review and searches of a variety of sources, such as completed assessments, 2016 and 2020 Chemical Data Reporting (U.S. EPA, 2020a, 2016). General population and environmental exposures were evaluated for the inhalation, dermal and ingestion exposure pathways based on environmental release data. In summary, modeled environmental release estimates were used as inputs for the general population exposure modeling.

6.4.1 General Population and Environmental Exposure: Surface, Groundwater, and Drinking Water

For the environmental exposure assessment, EPA used modeled surface water concentrations and sediment concentrations modeled via VVWM-PSC.

EPA conducted modeling with the U.S. EPA's Variable Volume Water Model with Point Source Calculator tool (VVWM-PSC), to estimate concentrations of BBP within surface water and sediment. VVWM-PSC considers model inputs of physical and chemical properties of BBP (*i.e.*, K_{OW}, K_{OC}, water column half-life, photolysis half-life, hydrolysis half-life, and benthic half-life) allowing EPA to model predicted surface water concentrations (U.S. EPA, 2019). The VVWM-PSC model was also used to estimate settled sediment in the benthic region of streams.

Where available, EPA compared reported environmental monitoring data and reported environmental modeling data with EPA modeled media concentrations. Section 4.2 of the *Environmental Media and General Population and Environment Exposure for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025m) summarizes measured concentrations of BBP within published literature for surface water, precipitation, and sediment. Section 4.1 of the *Environmental Media and General Population and Environment Exposure for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025m) presents modeled concentrations of BBP within surface water and sediment from surface water and wastewater for relevant COUs. Concentrations of BBP in surface water can lead to different exposure scenarios including dermal exposure [presented in Section 5.1.1 (U.S. EPA, 2025m)] or incidental ingestion exposure [Section 5.1.2 (U.S. EPA, 2025m)] to the general population swimming in affected waters. Exposure scenarios were assessed using the highest concentration of BBP in surface water based on highest releasing OES (Hydraulic Fluids). Additionally, modeled surface water concentrations were used to estimate drinking water exposures [Section 6 (U.S. EPA, 2025m)].

When applying the PSC, certain physicochemical parameters are used as model input variables, which are collected as a part of the fate team's assessment. The use of SR to verify physical and chemical properties of BBP are thus relevant for exposure modeling using the VVWM-PSC. Physical-chemical and fate properties selected by EPA for this assessment were applied as inputs to the PSC model and were sourced from parameters reviewed and described within the and *Physical Chemistry and Fate and Transport Assessment for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025p).

6.4.2 General Population and Environmental Exposure: Ambient Air

EPA evaluated general population and environmental exposures based on measured and predicted concentrations of BBP in ambient air. Section 8.1 and 8.2 of the *Environmental Media and General Population and Environment Exposure for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025m) summarizes the estimated ambient air concentrations and reported measured concentrations for ambient air found in the peer-reviewed from the systematic review, respectively. EPA estimated air releases were used as inputs for estimating ambient air concentrations and deposition fluxes via American Meteorological Society/Environmental Protection Agency Regulatory Model (AERMOD). A full description of input parameters is provided in Appendix B of the *Environmental Media and General Population and Environment Exposure for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025m). Modeled ambient air concentrations were used to estimate inhalation exposure. Modeled deposition fluxes were used to estimate soil concentrations of BBP in sections 8.3.1 of *Environmental Media and General Population and Environment Exposure for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025m). Where available, EPA compared reported environmental monitoring or systematic review data with AERMOD modeled ambient air concentrations.

6.4.3 General Population Exposure: Dietary, Biomonitoring and Exposure Reconstruction

Human milk and urinary biomonitoring data for BBP was collected through systematic review. BBP Biomonitoring data for human milk from the systematic review monitoring literature is summarized in Section 10.1 (Human Milk Exposures) of the *Environmental Media and General Population and Environment Exposure for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025m). EPA reviewed biomonitoring studies that measured BBP in human milk, and the highest measured concentration was used to screen for risks. The results supported EPA's decision to not quantitatively evaluate infant exposure to BBP via human milk ingestion.

BBP urinary biomonitoring data from the systematic review monitoring literature was considered. EPA relied on NHANES biomonitoring data analyzed in Section 10 of the *Environmental Media and General*

Population and Environment Exposure for Butyl Benzyl Phthalate (BBP) (U.S. EPA, 2025m). EPA focused on other agency risk evaluations to compare against EPA's own analysis of NHANES biomonitoring data.

6.4.4 Consumer Exposure Assessment

EPA assessed consumer exposure to BBP for both users and bystanders, resulting from use of consumer products and articles, see The *Consumer and Indoor Exposure Assessment for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025a). The major routes of exposure considered were via ingestion, inhalation, and dermal exposure. Consumer products containing BBP were identified through review and searches of a variety of sources, such as completed assessments, 2016 and 2020 Chemical Data Reporting (U.S. EPA, 2020a, 2016), in addition to chemical safety data sheets (SDSs) identified through product-specific internet searches. Chemical weight fractions were gathered from SDSs and completed assessments and used to tailor COU-specific consumer exposure scenarios for products and articles identified in the consumer market. The dermal assessment was based on DuPont (2006b) and DuPont (2006a), which was an *in vitro* absorption study using human skin.

6.4.4.1 Indoor Dust Monitoring

EPA evaluated consumer exposure to BBP through ingestion of indoor dust based on measured concentrations of BBP in representative residential scenarios. Section 4.1.2 of the *Consumer and Indoor Exposure Assessment for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025a) summarizes the indoor dust concentration data that was identified during systematic review. During systematic review, 89 studies containing potential indoor dust monitoring data for BBP were identified. Of these, 11 collected original data, were conducted in the U.S., reported high quality sampling and analytical methods, and measured dust in homes, offices, or other indoor environments that are representative of the U.S. general population. Out of the 11 studies, eight were selected because they collected settled indoor dust. Of these eight studies, only four reported the statistical summaries needed for this analysis, settled dust average and 95th percentile measured concentrations, which are used in the comparison to indoor dust ingestion modeling data. The measured data on BBP concentrations in residential indoor dust were used with dust intake rate estimates from Özkaynak et al. (2022) and body mass estimates from the Exposure Factors Handbook (U.S. EPA, 2011) to obtain an allometric estimate of BBP intake for consumers in residential household dust.

6.4.5 Other data sources

The exposure models relied heavily on the physical chemical and fate properties as input parameters. Sections 5.1 and 5.2 describe how the physical chemical and fate properties were selected. Where applicable, EPA relied on model defaults, exposure factors and activity patterns available from EPA's Exposure Factors Handbook (U.S. EPA, 2017). As mentioned previously, these physical chemical and fate parameters are used as inputs for PSC modeling of surface water concentrations of BBP and as inputs for AERMOD modeling.

6.5 Environmental and Human Health Hazard

Sections 7.4 and 7.5, the 2021 Draft Systematic Review Protocol explain how information from data sources that undergo systematic review and those that do not are considered for use in risk evaluations under TSCA, specifically, for evaluating environmental and human health hazard, respectively (U.S. EPA, 2021).

The environmental hazard evidence streams, as described in Table 7-8 of the 2021 Draft Systematic Review Protocol, have been updated to increase the level of clarity and consistency of granularity (U.S.

EPA, 2021). Table 6-1 the updated environmental hazard evidence streams that parses out the types of mechanistic data evidence streams.

6.5.1 Environmental Hazard

Section 7.4.1 of the 2021 Draft Systematic Review Protocol describes how environmental hazard integration is organized into different evidence streams. The environmental hazard evidence streams for risk evaluations conducted under TSCA, as described in Table 7-8 of the 2021 Draft Systematic Review Protocol, have been updated (Table 6-1; updates are represented in bold text) to increase the level of clarity and consistency of granularity (U.S. EPA, 2021). These updated environmental hazard evidence streams more clearly reflect how apical and mechanistic hazardous endpoints (as defined by the screening PECO statement) that result from either controlled field/laboratory or uncontrolled exposure field studies are binned to better consider the relevancy of the data for the respective risk evaluation.

Table 6-1. Querying the Evidence to Organize Integration for Environmental Data and Information

Evidence Stream	Questions
Apical endpoints (controlled field/laboratory conditions)	Of the available data, are there endpoints that could have population level effects such as reproduction, growth, and/or mortality?
Mechanistic data **(controlled field/laboratory conditions)**	Is the mechanistic endpoint linked to an apical endpoint? Is it part of an AOP? If not, can you instead use it qualitatively? If a transcriptomic point of departure (tPOD) is available, is it appropriate to use quantitatively?
Apical endpoints (uncontrolled exposure field conditions)	Are there any field studies available showing adverse effects? How does exposure to the chemical of interest affect the community of organisms? Are there any co-occurring adverse environmental conditions other than exposure to the chemical of interest that should be taken into consideration?
Mechanistic endpoints (uncontrolled exposure field conditions)	**Is the mechanistic endpoint linked to an apical endpoint? Is it part of an AOP? If not, can you instead use it qualitatively? If a transcriptomic point of departure (tPOD) is available, is it appropriate to use quantitatively? Are there any co-occurring adverse environmental conditions other than exposure to the chemical of interest that should be taken into consideration?**

As described in the *Environmental Hazard Assessment for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025l), streams for environmental hazard included empirical data with apical endpoints for aquatic and terrestrial organisms that were reviewed following the TSCA systematic review process.

EPA reviewed potential environmental health hazards associated with BBP (U.S. EPA, 2025l). Studies identified as meeting PECO screening criteria and evaluated for data quality received an overall quality determination of high, medium, low, or uninformative. Data on the toxicity of BBP were limited and

only high and medium-quality studies were used for purposes of hazard and risk characterization (U.S. EPA, 2025l).

Using empirical evidence, EPA characterized the environmental hazards of BBP to surrogate species representing various receptor groups (U.S. EPA, 2025l).

Evaluations of the strength of evidence and weight of scientific evidence for environmental hazard was conducted as described within Section 7.4.2 of the 2021 Draft Systematic Review Protocol (U.S. EPA, 2021). For additional details on the application of this methodology, please see Appendix B of the *Environmental Hazard Assessment for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025l).

6.5.2 Human Health Hazard

Section 7.5 of the 2021 Draft Systematic Review Protocol describes how EPA considers individual evidence streams (human, animal toxicity, and mechanistic/supplemental studies) when integrating evidence (U.S. EPA, 2021). For risk evaluations conducted under TSCA, the human health hazard evidence streams were updated (Table 6-2) to more clearly reflect how apical and mechanistic hazard endpoints (as defined by the screening PECO statement) that result from either animal toxicology and epidemiology studies are binned to better consider the relevancy of the data for the risk evaluation.

Table 6-2. Querying the Evidence to Organize Integration for Human Health Hazard Data and Information

Evidence Stream	Questions
Studies of Exposed Humans Considered for Deriving Toxicity Values	Is there any qualitative data in human studies that can be used to support PODs used for risk estimates?
In Vivo Mammalian Animal Studies Considered for Deriving Toxicity Values	Is there dose-response information and/or endpoints that could be used as PODs? Are there differences/similarities in toxicity across studies of different exposure durations and routes? Is there concordance across species and studies for observed endpoints?
Mechanistic and *In Vitro* Studies and Supplemental Information	Is the mechanistic endpoint linked to an apical endpoint? Is it part of an AOP? If not, can it be used qualitatively?

However, as discussed in Section 4.6.1 above, because of the wealth of existing assessments for BBP, a modified fit for purpose approach was employed. Rather than evaluating and integrating all evidence examining BBP exposure and all health outcomes, EPA focused on identifying studies that could inform an updated dose response assessment or supported identification of a new human health hazard. To do this, EPA first reviewed existing assessments of BBP (see Appendix B of the *Risk Evaluation of BBP* (U.S. EPA, 2025q)), which have consistently identified non-cancer liver, kidney and developmental

toxicity as the most sensitive non-cancer hazards associated with oral exposure to BBP in experimental animal models.

EPA decided that key studies used to support POD selection in existing assessments would also be important for its updated hazard and dose-response assessment of BBP. For purposes of this assessment, EPA considered key studies from existing assessments of BBP to be those considered for dose-response assessment and/or those used to establish a POD for subsequent use in risk characterization. Key studies were evaluated for data quality consistent with EPA's Draft Systematic Review Protocol (U.S. EPA, 2021). Because existing assessments of BBP have consistently identified liver, kidney, and developmental effects of BBP as the most sensitive effects, evidence streams were integrated for these health outcomes.

However, as further described in Section 4.6.1 above and in the *Non-Cancer Human Health Hazard Assessment for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025o), EPA also sought to identify new PECO-relevant literature published since the most recent existing assessment of BBP (EC/HC, 2015a). New PECO-relevant studies provided information pertaining to five health outcomes: reproductive/developmental, neurotoxicity, cardiovascular, immune system, and musculoskeletal. Therefore, evidence streams were also integrated for these non-cancer health outcomes, as well as for all cancer outcomes.

However, as further described in Section 4.6.1 above and in the *Non-Cancer Human Health Hazard Assessment for Butyl Benzyl Phthalate (BBP)* (U.S. EPA, 2025o), EPA also identified new reproductive/developmental toxicity studies published since the most recent existing assessments of BBP (EC/HC, 2015a; NICNAS, 2015). These studies met the PECO screening criteria and went through a further filtering step and were fully integrated (qualitative or quantitative) into the human health hazard assessment as part of the weight of scientific evidence for BBP.

As descried in Section 4.6.2.2, studies with HEDs within an order of magnitude of the lowest LOAEL-based HED identified across existing assessments were also considered sensitive and potentially relevant for POD selection. These studies were further reviewed by EPA to determine if they support a different human health hazard or potentially lower POD than those identified in existing assessments of BBP. Mechanistic studies and studies with HEDs more than an order of magnitude above the HEDs associated with the lowest LOAELs from previous assessments were integrated into the hazard identification and characterization process but did not undergo TSCA study quality evaluations. Instead, these studies were evaluated in a manner consistent with the Office of Pesticide Programs *Guidance for Considering and Using Open Literature Toxicity Studies to Support Human Health Hazard Assessment* (U.S. EPA, 2012). These studies were considered of sufficient quality to be considered qualitatively as part of the weight of scientific evidence and were assigned a quality score of medium.

6.6 Dermal Absorption

Table 6-3 describes relevant questions to consider when integrating evidence from empirical data, read-across analysis from analog chemicals, and models of dermal absorption.

For evaluating dermal exposures to BBP, EPA first considered available data related to the dermal absorption of BBP identified in Section 5.6. The dermal assessment was based on DuPont (2006b) and DuPont (2006a), which was an *in vitro* absorption study using human skin.

Table 6-3. Querying the Evidence to Organize Integration for Human Health Dermal Absorption

Evidence Stream (Individual or Combined)	Questions
Studies of Exposed Humans for the Target Chemical	Are there human studies that can be used quantitively to determine dermal absorption estimates or qualitatively in a weight of scientific evidence analysis?
In Vivo Mammalian Animal Studies for the Target Chemical	Are there *in vivo* animal data that can be used quantitatively or qualitatively?
In Vitro/Ex Vivo Studies and Supplemental Information for the Target Chemical	Are there *in vitro* dermal absorption data that can be used quantitatively or qualitatively?
Read Across From Chemical Analogs	Are there human, *in vivo*, or *in vitro/ex vivo* dermal absorption data available for analogs of the target chemical that have similar physical-chemical properties?
Models for Kp and Fraction Absorption	Are there models available to estimate the dermal permeability coefficient (Kp) or fraction absorbed?
Combining Evidence	Are there differences/similarities in dermal absorption across studies? Is there concordance within and across *in vivo* and *in vitro* studies as well as within and across species? If read-across analysis from an analog chemical is used, is there consistency with any limited data for the target chemical or among the analog chemical studies? If multiple models are used, is there concordance among the models and with any limited empirical data?

REFERENCES

Ahmad, R; Gautam, AK; Verma, Y; Sedha, S; Kumar, S. (2014). Effects of in utero di-butyl phthalate and butyl benzyl phthalate exposure on offspring development and male reproduction of rat. Environ Sci Pollut Res Int 21: 3156-3165. http://dx.doi.org/10.1007/s11356-013-2281-x

Ahmad, R; Verma, Y; Gautam, A; Kumar, S. (2015). Assessment of estrogenic potential of di-n-butyl phthalate and butyl benzyl phthalate in vivo. Toxicol Ind Health 31: 1296-1303. http://dx.doi.org/10.1177/0748233713491803

Alam, MS; Kurohmaru, M. (2015). Butylbenzyl phthalate induces spermatogenic cell apoptosis in prepubertal rats. Tissue Cell 48: 35-42. http://dx.doi.org/10.1016/j.tice.2015.12.001

BIBRA. (1986). Rat liver and lipid effects of representative phthalate esters with EPA acknowlegement letter [TSCA Submission]. (OTS0000478-0. FYI-AX-0286-0478. TSCATS/201638). Chemical Manufacturers Association. https://ntrl.ntis.gov/NTRL/dashboard/searchResults.xhtml?searchQuery=OTS00004780

Calafat, AM; Valentin-Blasini, L; Ye, X. (2015). Trends in Exposure to Chemicals in Personal Care and Consumer Products. Curr Environ Health Rep 2: 348-355. http://dx.doi.org/10.1007/s40572-015-0065-9

CDC, CfDCaP. (2022). Formaldehyde (FDH) (CAS RN: 50-00-0): NHANES Biomonitoring Data (Blood; Cigarette smoke). Atlanta, GA: Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/exposurereport/data_tables.html?NER_SectionItem=NHANES

Debartolo, D; Jayatilaka, S; Yan Siu, N; Rose, M; Ramos, RL; Betz, AJ. (2016). Perinatal exposure to benzyl butyl phthalate induces alterations in neuronal development/maturation protein expression, estrogen responses, and fear conditioning in rodents. Behav Pharmacol 27: 77-82. http://dx.doi.org/10.1097/FBP.0000000000000190

DuPont. (2006a). Polyvinyl chloride film plasticized with butyl benzyl phthalate: In vitro dermal absorption rate testing. (DuPont-17805). Independence, OH: Ferro Corporation, Inc.

DuPont. (2006b). [Sanitized] Butyl benzyl phthalate: In vitro dermal absorption rate testing. (DuPont-17755). Independence, OH: Ferro Corporation, Inc.

EC/HC. (2015a). State of the science report: Phthalate substance grouping 1,2-Benzenedicarboxylic acid, diisononyl ester; 1,2-Benzenedicarboxylic acid, di-C8-10-branched alkyl esters, C9-rich (Diisononyl Phthalate; DINP). Chemical Abstracts Service Registry Numbers: 28553-12-0 and 68515-48-0. Gatineau, Quebec. https://www.ec.gc.ca/ese-ees/default.asp?lang=En&n=47F58AA5-1

EC/HC. (2015b). State of the science report: Phthalate substance grouping: Medium-chain phthalate esters: Chemical Abstracts Service Registry Numbers: 84-61-7; 84-64-0; 84-69-5; 523-31-9; 5334-09-8;16883-83-3; 27215-22-1; 27987-25-3; 68515-40-2; 71888-89-6. Gatineau, Quebec: Environment Canada, Health Canada. https://www.ec.gc.ca/ese-ees/4D845198-761D-428B-A519-75481B25B3E5/SoS_Phthalates%20%28Medium-chain%29_EN.pdf

Fasano, WJ; Manning, LA; Green, JW. (2002). Rapid integrity assessment of rat and human epidermal membranes for in vitro dermal regulatory testing: correlation of electrical resistance with tritiated water permeability. Toxicol In Vitro 16: 731-740. http://dx.doi.org/10.1016/s0887-2333(02)00084-x

Gray, LE; Lambright, CS; Conley, JM; Evans, N; Furr, JR; Hannas, BR; Wilson, VS; Sampson, H; Foster, PMD. (2021). Genomic and Hormonal Biomarkers of Phthalate-Induced Male Rat Reproductive Developmental Toxicity Part II: A Targeted RT-qPCR Array Approach That Defines a Unique Adverse Outcome Pathway. Toxicol Sci 182: 195-214. http://dx.doi.org/10.1093/toxsci/kfab053

Health Canada. (2020). Screening assessment - Phthalate substance grouping. (En14-393/2019E-PDF). Environment and Climate Change Canada, Health Canada. https://www.canada.ca/en/environment-climate-change/services/evaluating-existing-

substances/screening-assessment-phthalate-substance-grouping.html
Howdeshell, KL; Wilson, VS; Furr, J; Lambright, CR; Rider, CV; Blystone, CR; Hotchkiss, AK; Gray, LE, Jr. (2008). A mixture of five phthalate esters inhibits fetal testicular testosterone production in the Sprague-Dawley rat in a cumulative, dose-additive manner. Toxicol Sci 105: 153-165. http://dx.doi.org/10.1093/toxsci/kfn077
ILS. (2017). Pubertal development and thyroid function in intact juvenile/peripubertal female rats; benzyl butyl phthalate. (10005.0302). Research Triangle Park, NC: RTI International.
Jahreis, S; Trump, S; Bauer, M; Bauer, T; Thürmann, L; Feltens, R; Wang, Q; Gu, L; Grützmann, K; Röder, S; Averbeck, M; Weichenhan, D; Plass, C; Sack, U; Borte, M; Dubourg, V; Schüürmann, G; Simon, JC; Martin Von, B; Hackermüller, J; Eils, R; Lehmann, I; Polte, T. (2018). Maternal phthalate exposure promotes allergic airway inflammation over 2 generations through epigenetic modifications. J Allergy Clin Immunol 141: 741-753. http://dx.doi.org/10.1016/j.jaci.2017.03.017
Kwack, S; Kim, K; Kim, H; Lee, B. (2009). Comparative toxicological evaluation of phthalate diesters and metabolites in Sprague-Dawley male rats for risk assessment. J Toxicol Environ Health A 72: 1446-1454. http://dx.doi.org/10.1080/15287390903212923
Lee, BM; Koo, HJ. (2007). Hershberger assay for antiandrogenic effects of phthalates. J Toxicol Environ Health A 70: 1365-1370. http://dx.doi.org/10.1080/15287390701432285
Min, A; Liu, F; Yang, X; Chen, M. (2014). Benzyl butyl phthalate exposure impairs learning and memory and attenuates neurotransmission and CREB phosphorylation in mice. Food Chem Toxicol 71: 81-89. http://dx.doi.org/10.1016/j.fct.2014.05.021
Nagao, T; Ohta, R; Marumo, H; Shindo, T; Yoshimura, S; Ono, H. (2000). Effect of butyl benzyl phthalate in Sprague-Dawley rats after gavage administration: A two-generation reproductive study. Reprod Toxicol 14: 513-532. http://dx.doi.org/10.1016/S0890-6238(00)00105-2
Nakagomi, M; Suzuki, E; Saito, Y; Nagao, T. (2017). Endocrine disrupting chemicals, 4-nonylphenol, bisphenol A and butyl benzyl phthalate, impair metabolism of estradiol in male and female rats as assessed by levels of 15α-hydroxyestrogens and catechol estrogens in urine. J Appl Toxicol 38: 688-695. http://dx.doi.org/10.1002/jat.3574
NICNAS. (2015). Priority existing chemical assessment report no. 40: Butyl benzyl phthalate. (PEC40). Sydney, Australia: Australian Government Department of Health and Ageing. https://heronet.epa.gov/heronet/index.cfm/reference/download/reference_id/3664467
OECD. (2004a). Guidance document for the conduct of skin absorption studies. (ENV/JM/MONO(2004)2). https://www.oecd-ilibrary.org/environment/guidance-document-for-the-conduct-of-skin-absorption-studies_9789264078796-en;jsessionid=R95G2RBh7UhKBExKA28kPPHPPmxc8LYbY3f6KXt8.ip-10-240-5-151
OECD. (2004b). Test No. 427: Skin absorption: in vivo method. Paris, France.
OECD. (2004c). Test No. 428: Skin absorption: In vitro method. Paris, France. http://dx.doi.org/10.1787/9789264071087-en
OECD. (2011a). Emission scenario document on coating application via spray-painting in the automotive refinishing industry. In OECD Series on Emission Scenario Documents No 11. (ENV/JM/MONO(2004)22/REV1). Paris, France: Organization for Economic Co-operation and Development. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2004)22/rev1&doclanguage=en
OECD. (2011b). Guidance notes on dermal absorption. (ENV/JM/MONO(2011)36). https://www.oecd.org/env/ehs/testing/48532204.pdf
OECD. (2022). Series on Testing & Assessment, No. 156: Guidance notes on dermal absorption studies (Second edition). (ENV/JM/MONO(2011)36/REV1). Paris, France: Organisation for Economic Co-operation and Development (OECD).

https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV-JM-MONO(2011)36%20&doclanguage=en
Özkaynak, H; Glen, G; Cohen, J; Hubbard, H; Thomas, K; Phillips, L; Tulve, N. (2022). Model based prediction of age-specific soil and dust ingestion rates for children. J Expo Sci Environ Epidemiol 32: 472-480. http://dx.doi.org/10.1038/s41370-021-00406-5
QuanTech. (2021). American Healthy Homes Survey, final report: Data documentation. (AHHSII). Washington, DC: U.S. Department of Housing and Urban Development. https://www.hud.gov/program_offices/healthy_homes
Radke, EG; Braun, JM; Meeker, JD; Cooper, GS. (2018). Phthalate exposure and male reproductive outcomes: A systematic review of the human epidemiological evidence [Review]. Environ Int 121: 764-793. http://dx.doi.org/10.1016/j.envint.2018.07.029
Radke, EG; Braun, JM; Nachman, RM; Cooper, GS. (2020). Phthalate exposure and neurodevelopment: A systematic review and meta-analysis of human epidemiological evidence [Review]. Environ Int 137: 105408. http://dx.doi.org/10.1016/j.envint.2019.105408
Schmitt, EE; Vellers, HL; Porter, WW; Lightfoot, JT. (2016). Environmental endocrine disruptor affects voluntary physical activity in mice. Med Sci Sports Exerc 48: 1251-1258. http://dx.doi.org/10.1249/MSS.0000000000000908
TNO CIVO. (1993). Dietary one-generation reproduction study with butyl benzyl phthalate in rats with cover letter dated 040793 [TSCA Submission]. (EPA/OTS Doc #86-930000189). St. Louis, MO: Monsanto Co. https://ntrl.ntis.gov/NTRL/dashboard/searchResults.xhtml?searchQuery=OTS0538169
Tyl, RW; Myers, CB; Marr, MC; Fail, PA; Seely, JC; Brine, DR; Barter, RA; Butala, JH. (2004). Reproductive toxicity evaluation of dietary butyl benzyl phthalate (BBP) in rats. Reprod Toxicol 18: 241-264. http://dx.doi.org/10.1016/j.reprotox.2003.10.006
U.S. EPA. (2011). Exposure Factors Handbook, Chapter 8: Body weight studies. Washington, DC. https://www.epa.gov/expobox/exposure-factors-handbook-chapter-8
U.S. EPA. (2012). Guidance for considering and using open literature toxicity studies to support human health risk assessment. Washington, DC: U.S. Environmental Protection Agency, Office of Pesticide Programs. https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/guidance-considering-and-using-open-literature
U.S. EPA. (2016). Chemical data reporting: 2016 data. Washington, DC: U.S. Environmental Protection Agency, Chemical Data Reporting. Retrieved from https://www.epa.gov/chemical-data-reporting/access-cdr-data#2016
U.S. EPA. (2017). Update for Chapter 5 of the Exposure Factors Handbook: Soil and dust ingestion [EPA Report]. (EPA/600R-17/384F). Washington, DC: National Center for Environmental Assessment, Office of Research and Development. https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P100TTX4.txt
U.S. EPA. (2018). Application of systematic review in TSCA risk evaluations. (740-P1-8001). Washington, DC: U.S. Environmental Protection Agency, Office of Chemical Safety and Pollution Prevention. https://www.epa.gov/sites/production/files/2018-06/documents/final_application_of_sr_in_tsca_05-31-18.pdf
U.S. EPA. (2019). Point Source Calculator: A Model for Estimating Chemical Concentration in Water Bodies. Washington, DC: U.S. Environmental Protection Agency, Office of Chemical Safety and Pollution Prevention.
U.S. EPA. (2020a). 2020 CDR data [Database]. Washington, DC: U.S. Environmental Protection Agency, Office of Pollution Prevention and Toxics. https://www.epa.gov/chemical-data-reporting/access-cdr-data
U.S. EPA. (2020b). Final scope of the risk evaluation for butyl benzyl phthalate (1,2-benzenedicarboxylic acid, 1-butyl 2-(phenylmethyl) ester); CASRN 85-68-7 [EPA Report].

(EPA-740-R-20-015). Washington, DC: Office of Chemical Safety and Pollution Prevention. https://www.epa.gov/sites/default/files/2020-09/documents/casrn_85-68-7_butyl_benzyl_phthalate_finalscope.pdf

U.S. EPA. (2021). Draft systematic review protocol supporting TSCA risk evaluations for chemical substances, Version 1.0: A generic TSCA systematic review protocol with chemical-specific methodologies. (EPA Document #EPA-D-20-031). Washington, DC: Office of Chemical Safety and Pollution Prevention. https://www.regulations.gov/document/EPA-HQ-OPPT-2021-0414-0005

U.S. EPA. (2022a). Ambient Monitoring Technology Information Center (AMTIC) - Ambient Monitoring Archive for HAPs [Database]. Washington, DC. Retrieved from https://www.epa.gov/amtic/amtic-ambient-monitoring-archive-haps

U.S. EPA. (2022b). Formaldehyde (FDH) (CAS RN: 50-00-0): WQP Output (NWIS, STEWARDS & STORET), Site data & sample results (physical/chemical metadata) [Database]: National Water Quality Monitoring Council. Retrieved from https://acwi.gov/monitoring/waterqualitydata.html

U.S. EPA. (2025a). Consumer and Indoor Dust Exposure Assessment for Butyl benzyl phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025b). Data Extraction Information for Environmental Hazard and Human Health Hazard Animal Toxicology and Epidemiology for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025c). Data Extraction Information for General Population, Consumer, and Environmental Exposure for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025d). Data Quality Evaluation and Data Extraction Information for Dermal Absorption for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025e). Data Quality Evaluation and Data Extraction Information for Environmental Fate and Transport for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025f). Data Quality Evaluation and Data Extraction Information for Environmental Release and Occupational Exposure for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025g). Data Quality Evaluation and Data Extraction Information for Physical and Chemical Properties for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025h). Data Quality Evaluation Information for Environmental Hazard for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025i). Data Quality Evaluation Information for General Population, Consumer, and Environmental Exposure for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025j). Data Quality Evaluation Information for Human Health Hazard Animal Toxicology for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025k). Data Quality Evaluation Information for Human Health Hazard Epidemiology for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025l). Environmental Hazard Assessment for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025m). Environmental Media and General Population and Environment Exposure for Butyl benzyl phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025n). Environmental Release and Occupational Exposure Assessment for Butyl benzyl phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.

U.S. EPA. (2025o). Non-cancer Human Health Hazard Assessment for Butyl benzyl phthalate (BBP).

Washington, DC: Office of Pollution Prevention and Toxics.
U.S. EPA. (2025p). Physical Chemistry and Fate and Transport Assessment for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.
U.S. EPA. (2025q). Risk Evaluation for Butyl Benzyl Phthalate (BBP). Washington, DC: Office of Pollution Prevention and Toxics.