Comments from the Reproductive Toxicology and Epidemiology Section, OEHHA, Cal/EPA, on the Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-Based Health Assessments

Step 1: “Prepare Topic” – Formulate PECO question(s) and develop draft protocol

Comments on Step 1:

- This appears to be an appropriate approach to narrowly focused toxicological questions (e.g. “Do women exposed to chemical X in non-occupational settings have reduced fertility?).
- On the other hand, the PECO and protocol approach would have limited applicability to broader public health questions (e.g. “What is the most sensitive toxicological endpoint for this chemical?”).
- It would be helpful if OHAT would clarify whether it endorses the application of PECO and protocol development to regulatory hazard identification in general.
- Legislative requirements and regulatory guidance can impose broad public-health-based questions on government risk assessors, restricting potential for scoping and defining.

Step 2: Document details of literature search for reproducibility, and screen references for inclusion based on detailed criteria outlined in protocol.

Comments on Step 2:

Setting guidance for search and inclusion procedures, along with careful documentation of procedures, would be good practice. However, a few issues arise:

- While a detailed search *process* should be reproducible, the results of that process will change with updates and alterations to databases.
- The OHAT procedure emphasizes the process by which references are obtained.
  - Instead, emphasis should be on the results or goal of obtaining a complete and comprehensive set of relevant references.
  - The goal of obtaining a complete and comprehensive set of relevant references may be achievable by a variety of means.

With regards to selection criteria:

- Selection criteria are to be established prior to review of the scientific literature. However, establishing appropriate criteria would seem to require at least some

level of previous familiarity with the available scientific literature on the topic under consideration.
- Application of scientific judgment is intentionally minimized in the process. However, determining the “applicable outcomes, relevant exposures, and types of studies,” that should be included in the review are unlikely to be fully understood without some application of scientific judgment.

Step 3: Extract data from individual studies selected for inclusion using separate template forms for human, animal, and *in vitro* studies that are customized as needed for specific evaluations.

Comments on Step 3:

This step describes a good approach to documenting the elements that currently comprise any well-conducted evaluation process. Difficulties may arise in devising, using, or adapting a standardized data extraction form to capture data from studies having complex protocols. Some multigenerational animal studies, for example, include interim evaluations, extraction of subgroups for continuation in alternate experiments, etc.

Step 4: Assess the quality of individual studies by using a set of questions to evaluate study design and performance.

- **Five properties — risk of bias, unexplained inconsistency, indirectness, imprecision, and publication bias — are considered in downgrading an initial confidence rating.**
- **Four properties — large magnitude of effect, dose-response, all plausible confounding, and cross species/population/study consistency — are considered in upgrading an initial confidence rating.**

Comments on Step 4:

While having clear criteria for evaluating study quality is of value, there is a risk that good quality studies will be found to have limitations specific to the OHAT process, and so be explicitly or effectively downgraded relative to other studies. In particular, academic articles published in peer reviewed journals may be at a disadvantage as compared to large scale, GLP industry studies submitted to government agencies for regulatory purposes.

- Space limitations imposed by journal publishers can limit reporting of methodology and results for all measured outcomes.
- Animal toxicology studies, for example, do not necessarily blind researchers to treatment group throughout the study. While blind data collection may be ideal, it should be acknowledged that endpoints are not equally susceptible to this kind of bias.
    - Subjective measures such as clinical diagnosis, histological descriptions, or some behavioral observations are more easily affected by observer bias.
    - Objective measures such as body weight, numbers of offspring, cell counts, or machine-collected data are much less sensitive to observer expectations.

Other potentially useful tools and criteria to consider in evaluating study quality have not been explored in the draft document. For example:

- Citation analysis could be helpful in determining how a particular study is regarded.
- Quality of the publishing journal should be taken into account.
- Account should be taken of papers originating from the same lab and/or researchers, as compared to work originating from different labs. Multiple publications from the same group may not be entirely independent, and should be evaluated accordingly.

Step 5: Rate the Confidence in the Body of Evidence

- **Five properties — overall risk of bias, unexplained inconsistency, indirectness, imprecision, and publication bias — are considered to determine if the initial confidence rating should be downgraded.**
- **Four properties — large magnitude of effect, all plausible confounding, cross-species/population/study consistency, other (e.g. rare outcomes) — are considered in determining if the initial confidence rating should be upgraded.**

Comments on Step 5:

- With regards to risk of bias assessment, concerns would be the same as those expressed above for assessing bias in individual studies.
- What is meant by "unexplained" inconsistency? For example, concordance in results of developmental toxicity studies is not presumed across species. Investigators may or may not understand the mechanism underlying particular species differences. Would such a species difference then be considered "explained?"
- "Indirectness" shouldn't necessarily reduce confidence in the body of evidence. For example, endpoints of developmental toxicity observed in animals are not expected to be necessarily the same as what would be observed in humans. Under U.S. EPA's Risk Assessment Guidelines for Developmental Toxicity, "Evidence of a biologically significant increase in any of the four manifestations of developmental toxicity is considered indicative of an agent's potential for disrupting development and producing a developmental hazard."
- Magnitude could be an intrinsic property of a particular endpoint, but not reflect its biological importance as a toxicological effect — or its statistical detectability.
- Rare outcomes, such as findings of unusual birth defects showing a dose relationship but not necessarily statistical significance, should definitely be taken into consideration.

Step 6: Translate Confidence Ratings into Level of Evidence for Health Effect

Comments on Step 6:

Discussion of the descriptors to be used in describing the confidence in the database indicating that exposure is (or is not) associated with the health effect is clear and reasonable.

Step 7: Integrate Evidence to Develop Hazard Identification Conclusions

Comments on Step 7:

Discussion of the hazard identification conclusion categories, integration of the evidence streams for human and non-human animal studies, and incorporation of other relevant data is clear and reasonable.

However, delaying consideration of other relevant data, which "could include, but are not limited to, mechanistic data, *in vitro* data, or data based on upstream indicators of a health effect," to the final step of the process raises some concerns.

- "Other relevant data" are discussed only as support or opposition to the biological plausibility of a cause and effect relationship between exposure and an adverse health effect – and relegated to use in upgrading or downgrading hazard identification.

- Biological plausibility (e.g. temporal relationship of exposure and effect) should be considered at each stage of evaluating individual studies as well as the overall data set, and not just in making final level-of-hazard judgments.
- Additionally, mechanistic data, *in vitro* studies, and upstream indicators can provide direct evidence of a compound's toxic potential, and should be considered for that purpose.