Integrated Environmental Assessment and Management — Volume 13, Number 4—pp. 573–579
Received: 17 February 2017 | Returned for Revision: 20 February 2017 | Accepted: 4 April 2017

Brief Communication

New Approach to Weight-of-Evidence Assessment of Ecotoxicological Effects in Regulatory Decision-Making

A Tilghman Hall,† Scott E Belanger,‡ Pat D Guiney,§ Malyka Galay-Burgos,‖ Gerd Maack,# William Stubblefield,†† and Olwenn Martin‡‡*

†*Bayer AG, Monheim am Rhein, Germany*
‡*Procter & Gamble, Cincinnati, Ohio, USA*
§*University of Wisconsin, Madison, Wisconsin, USA*
‖*European Centre for Environmental Toxicology and Chemistry, Brussels, Belgium*
#*German Environmental Protection Agency (UBA), Dessau-Roßlau, Germany*
††*Oregon State University, Corvallis, Oregon, USA*
‡‡*Brunel University London, Uxbridge, United Kingdom*

EDITOR'S NOTE:

This paper represents 1 of 4 companion articles resulting from a SETAC Pellston Workshop® on "Improving the Usability of Ecotoxicology in Regulatory Decision-Making," held August 2015 in Shepherdstown, West Virginia, USA. The main workshop objectives were to improve the reliability and reproducibility of ecotoxicity studies, improve the use of peer-reviewed studies in regulatory risk assessment of chemicals, and improve the methods used in risk assessments when evaluating single or multiple lines of evidence.

ABSTRACT

Ecological risk assessments and risk management decisions are only as sound as the underlying information and processes to integrate them. It is important to develop transparent and reproducible procedures a priori to integrate often-heterogeneous evidence. Current weight-of-evidence (WoE) approaches for effects or hazard assessment tend to conflate aspects of the assessment of the quality of the data with the strength of the body of evidence as a whole. We take forward recent developments in the critical appraisal of the reliability and relevance of individual ecotoxicological studies as part of the effect or hazard assessment of prospective risk assessments and propose a streamlined WoE approach. The aim is to avoid overlap and double accounting of criteria used in reliability and relevance with that used in current WoE methods. The protection goals, problem formulation, and evaluation process need to be clarified at the outset. The data are first integrated according to lines of evidence (LoEs), typically mechanistic insights (e.g., cellular, subcellular, genomic), in vivo experiments, and higher-tiered field or observational studies. Data are then plotted on the basis of both relevance and reliability scores or categories. This graphical approach provides a means to visually assess and communicate the credibility (reliability and relevance of available individual studies), quantity, diversity, and consistency of the evidence. In addition, the external coherence of the body of evidence needs to be considered. The final step in the process is to derive an expression of the confidence in the conclusions of integrating the information considering these 5 aspects in the context of remaining uncertainties. We suggest that this streamlined approach to WoE for the effects or hazard characterization should facilitate reproducible and transparent assessments of data across different regulatory requirements. *Integr Environ Assess Manage* 2017;13:573–579. © 2017 The Authors. *Integrated Environmental Assessment and Management* published by Wiley Periodicals, Inc. on behalf of Society of Environmental Toxicology & Chemistry (SETAC)

Keywords: Effects characterization Hazard characterization Weight-of-evidence Relevance Reliability

This article includes online-only Supplemental Data.

* Address correspondence to Tilghman.hall@bayer.com

Published 17 March 2017 on wileyonlinelibrary.com/journal/ieam.

INTRODUCTION

In ecological risk assessment (ERA), it is necessary at the various stages of hazard or effect, exposure, and risk characterization to evaluate the available evidence, assess the fit of evidence within the problem formulation, and determine the extent to which conclusions can be supported by the overall body of evidence. For a risk management

decision to be supported and accepted, this process of integrating evidence should be objective, transparent, well documented, usable, resource efficient, reproducible, and understandable. Although the term has been used widely with differing meanings, this process is generally referred to as "weight-of-evidence" (WoE). Weight-of-evidence should be flexible enough to adjust to regulatory needs, deal with new data and information from multiple sources, and recognize that expert judgment plays a key role in the process.

It is important to stress that integrating evidence for hazard or effect assessment is not an end in itself but rather fits within a wider risk assessment and management paradigm. Figure 1 describes how such an effects characterization fits in the ERA framework. Depending on the regulatory context and the associated problem formulation plan, the outcome of the WoE exercise can be flexible and sufficiently adaptable to do things such as:

1) select a key study for a prospective ERA,
2) select studies for individual species from which a species sensitivity distribution may be drawn to derive an environmental quality standard,
3) select the most appropriate lines of evidence (LoEs; e.g., molecular, in vivo, higher-tiered field studies) to address a specific risk hypothesis, or
4) assess the probability of causality in a retrospective ERA.

As noted by Suter and Cormier (2011), the diversity of potential applications is too great and the preferences of individual risk assessors, decision makers, and stakeholders are simply too varied to allow for a dictated WoE method. A formal declaration of the explicit method that is being used in the WoE process is therefore of great importance (Suter and Cormier 2011).

For a prospective risk assessment, the regulatory framework and related problem formulation will greatly influence the WoE approach. A logical initial step in an effects characterization consists of aggregating evaluations of the reliability and relevance of individual studies. Thorough reviews of recent insights and developments in the evaluation of the reliability and relevance of ecotoxicological (effects) data can be found in the companion papers of Moermond et al. (this issue) and Rudén et al. (this issue), respectively. We describe here a simplified WoE approach that builds on this recent work and focuses specifically on the integration and evaluation of the overall body of evidence. The evaluation of the reliability and relevance of individual studies has traditionally been considered as part of the WoE process. To avoid overlap and double accounting, we propose here a streamlined strategy to integrate outcomes from several LoEs. When conducted cooperatively with all stakeholders, we suggest this improved WoE process will lead to more transparent risk assessment and management decisions, identifying remaining uncertainties and any additional testing needs.

Relevance and reliability in effects characterization

Reliability and relevance evaluations of individual studies provide the basis for an effects or hazard assessment WoE framework (see Figure 1). The approach is not intended to be overly prescriptive; therefore such evaluations should be adjusted to fit the protection goals required by the appropriate regulation being considered.

Reliability criteria. Moermond et al. (this issue) reviewed some of the different approaches used to define reliability. The working definition for reliability is "the inherent quality of an effect value in a test report or publication relating to: (1) a clearly described experimental design to allow for the study to be repeated independently, (2) the way the experimental

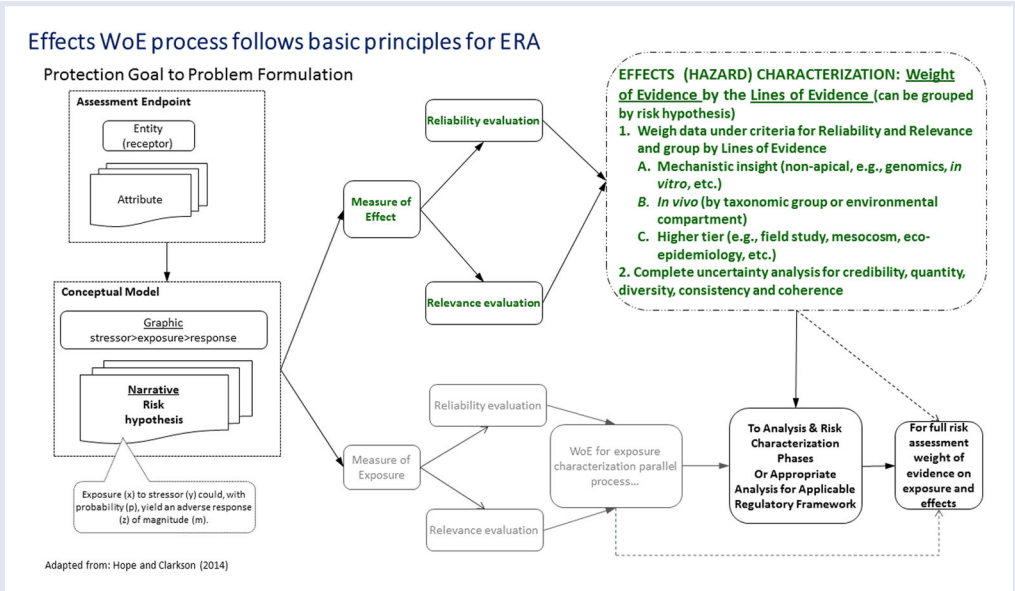

Figure 1. Environmental Risk Assessment (ERA) framework that integrates the weight-of-evidence (WoE) approach within the effects (hazard) characterization and uses relevance and reliability criteria as the basis of the analysis. Adapted from Hope and Clarkson (2014).

procedure were performed, and (3) the reporting of the results to provide evidence of the reproducibility and accuracy of the findings."

Relevance criteria. The key points regarding relevance were summarized by Rudén et al. (this issue), noting that a relevance assessment of ecotoxicity studies is required to ensure appropriate use in hazard and risk assessment of chemicals. Relevance assessment can be divided into 3 categories: regulatory relevance (fit for purpose to the regulatory framework, protection goal, and assessment endpoints), biological relevance (e.g., related to the test species, life stage, endpoints, and response function), and exposure relevance (e.g., related to test substance, exposure route and exposure dynamics).

Integrating relevance and reliability together for an effects characterization

Understanding the credibility of the whole body of evidence as a result of both the reliability and the relevance of individual ecotoxicological studies is an important concept in the WoE assessment. Studies that are found to be both reliable and relevant should carry the greatest weight in the risk assessment. Highly relevant and reliable studies may be used as the primary information, and those studies with mixed scores (medium relevance and reliability) provide corroborative support, whereas studies with low scores for both reliability and relevance should likely only be used qualitatively rather than quantitatively, or not used at all. This will depend on both the regulatory context and the amount of data available. To this end, we propose to present the body of evidence visually using either a matrix for categorical scores or a graph if continuous values are derived when assessing the reliability and relevance of individual studies. Specifically, each study is plotted within the line of evidence to which it belongs according to its relevance as the abscissa and reliability as the ordinate (Figure 2). By examining the suite of data available, one may be able to easily discern trends in the data and potentially sensitive taxonomic groups. No matter the purpose or the amount of data, all the information or LoEs need to be considered and shown for transparency.

Approaches should be accurately described. In the supplemental information (Supplemental Data 1), for illustrative purposes, we have taken the criteria discussed by Moermond et al. (this issue) and grouped them into an overall reliability assessment for study fitness-for-purpose, protocol repeatability, test replicability, statistical methods, and data reporting. Similarly, we have taken the relevance criteria and grouped them into 5 categories, balancing the 5 reliability categories to facilitate graphical presentation of the data. It is important to note here that the selection of 5 categories within each reliability and relevance criterion is simply an example and can be easily adjusted to fit the specific needs of the intended WoE approach for the regulatory framework.

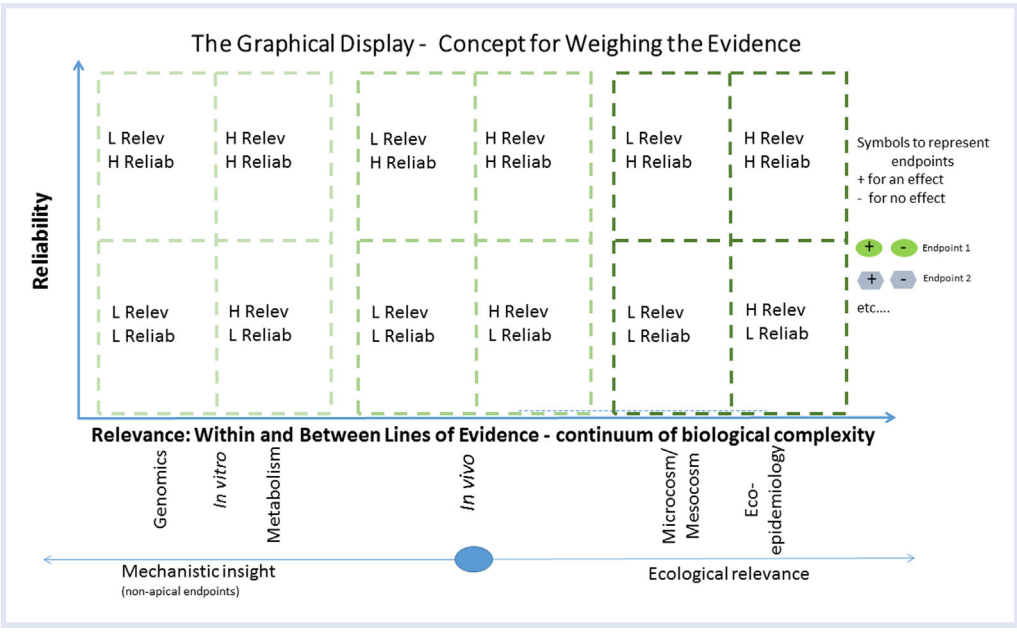

Figure 2. Representative weight-of-evidence (WoE) matrix using 3 lines of evidence (LoEs) that may be populated by studies with varying degrees of reliability and relevance. Typically, in vivo (whole organism) studies act as phenotypic anchors in environmental risk assessment. LoEs at lower levels (cellular, subcellular, genomic, etc.) of biological organization provide mechanistic insight that is valuable to understand in vivo responses. As one proceeds to levels of higher biological complexity, ecological relevance increases. For purposes of illustration, reliability and relevance can be plotted within the appropriate LoE matrix, each symbol representing a different taxonomic group (e.g., alga, invertebrate, or fish species) or type of measurement informing the LoE (e.g., gill cytoxicity in the mechanistic realm and a macroinvertebrate community principal response curve in the mesocosm realm). Endpoints common to different groups (e.g., *Daphnia magna* survival or fish survival) share a common color code. Each study endpoint has a symbol and, is graded with a plus (+) or minus (–) at this level to indicate if an effect or no effect was observed, respectively.

 DOI: 10.1002/ieam.1936

Examples of the graphical outcome of this approach are illustrated in Figure 3 with a hypothetical data set comprised of 14 studies (2 algal, 5 zooplankton, 3 insect or invertebrate, and 3 fish supplemented by 1 in vitro study). In the first example, the relevance and reliability of individual studies are evaluated in a regulatory context where expected environmental concentrations are not directly relevant (e.g., effects or hazard assessment needed as a basis for environmental quality standards) (Figure 3A). It appears that insects are the most relevant taxa, but these studies were attributed low reliability because exposures were not analytically confirmed. Two of five zooplankton studies were judged to be highly reliable because an experimentally robust test design was used as opposed to another two that lacked analytical confirmation. The fifth zooplankton study on reproduction was deemed somewhat relevant but of low reliability. The relevance of the fish studies is moderate because these were the results of short-term juvenile growth studies only, compared to life-cycle experiments of the insects studied and the zooplankton experiments. The results displayed in Figure 3A show a range and potential sensitivity relative to the different taxa.

In another regulatory context, an understanding of the range of environmental concentrations may be of relevance (e.g., selection of a key study or LoE to gain insights into a specific risk hypothesis). In Figure 3B, the environmental exposure range in the environment is known and can then be considered when evaluating relevance. This illustrates that the relevance of individual studies is dependent on the problem formulation whereas reliability is an intrinsic

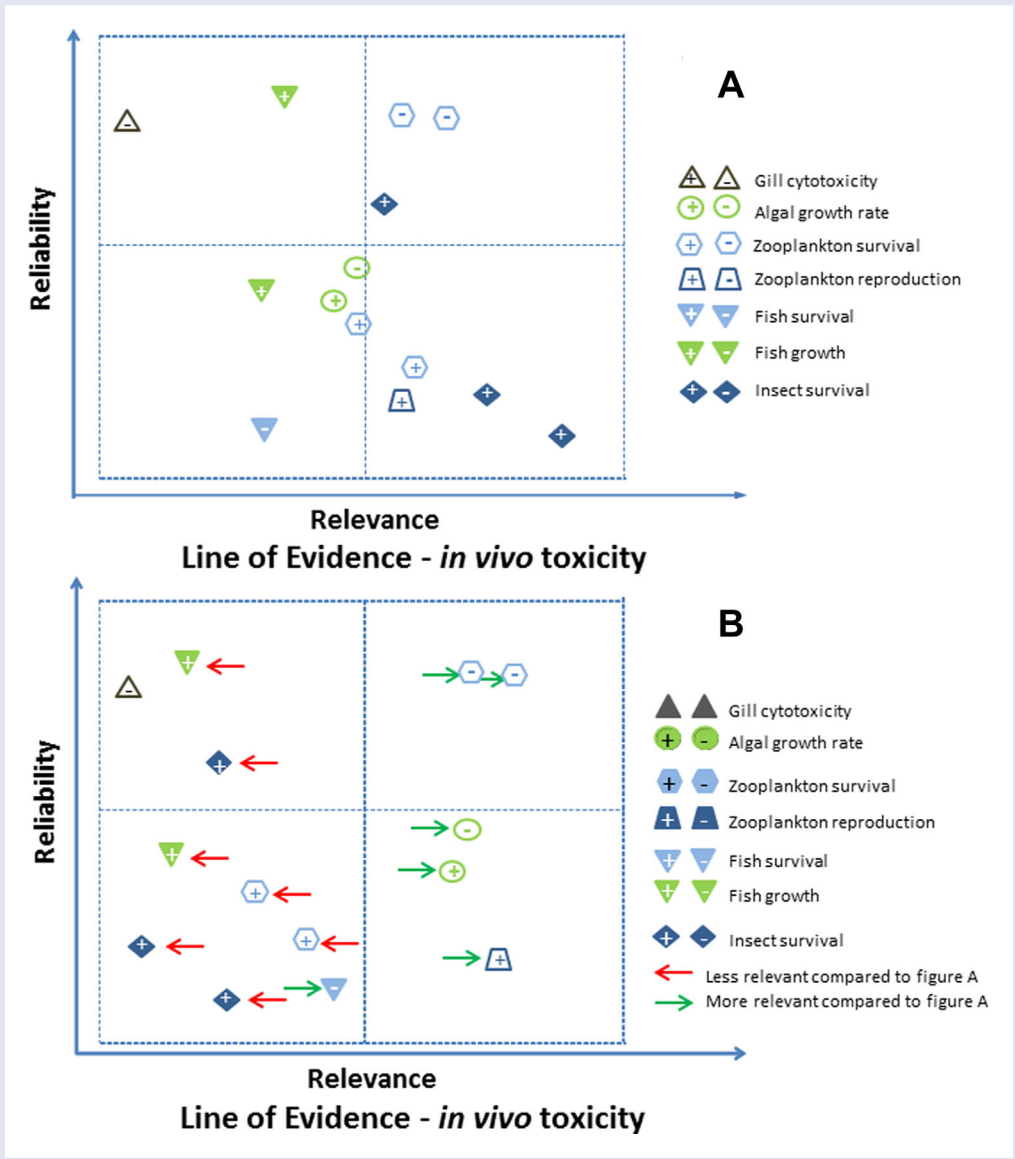

Figure 3. Hypothetical example applying the weight-of-evidence (WoE) matrix to an in vivo sample data set with and without considering a relevant environmental concentration: the data evaluation for a prospective risk assessment (A), and the changes in data evaluation if defined environmental exposure concentration is used as one of the primary criteria for relevance (B). The arrows indicate the direction for which relevance may shift for each study, given the environmental concentrations.

property of studies and scores do not change. For example, in this hypothetical case, the insect studies, judged as relevant for the purpose in Panel A are found to be exposed at unrealistically high concentrations thereby lowering their relevance considerably. Simultaneously, relevance for the algal studies and the zooplankton studies displaying no effects in survival was increased because these studies were performed in the range of plausible exposures. On the other hand, the zooplankton studies that displayed survival effects were also conducted with concentrations higher than those found in the environment, making these studies less relevant, similar to the 2 fish studies that displayed growth effects, making them even less relevant. The fish study that did not find an effect, however, was conducted at lower, more environmentally realistic concentrations, making this study more relevant, but still less relevant compared to the full or even multigeneration algae and zooplankton studies.

Relationship between reliability and relevance to WoE and uncertainty analysis

Cormier et al. (2010) suggested 4 considerations (typically applied as retrospective criteria) to assist judgments of studies into a WoE. These are credibility, coherence, strength, and diversity. A comparison of the considerations of Cormier et al. (2010) with those of Bradford Hill (1965) criteria are provided in Supplemental Data 2 and also illustrate how they are accounted for either within the critical appraisal of the reliability and relevance of individual studies or when considering the body of evidence as a whole within our framework. Plotting reliability against relevance (as shown in Figure 2) allows one to readily visualize and communicate most of these concepts.

The credibility of the body of evidence is related to the relevance and reliability of the entire body of evidence as a result of the relevance and reliability of individual studies (and across taxa and LoEs). Coherence according to Cormier et al (2010) considered consistency within an LoE and coherence with general knowledge together. We propose to consider both of these concepts separately and in turn. Consistency within an LoE may include exposure, the endpoint or type of effect observed, and the magnitude of the response or, if available, the dose-response relationships in the data. Identification of sensitive or vulnerable groups, relationships between acute and chronic effects, and presence or absence of thresholds should be noted. Inconsistencies should be identified and addressed. Coherence between LoEs and with other general knowledge indicates the extent to which the body of evidence is externally consistent with our current scientific understanding (e.g., mode or mechanism of action) and whether the findings are logically explained by known facts.

Strength of the body of evidence includes the amount of information available for each LoE (considered here under our quantity criterion), its relevance to the risk assessment hypotheses (accounted under credibility), as well as whether pieces of evidence are logically compelling and/or present quantitatively strong relationships (both already considered under consistency). In addition, the diversity of sources and types of evidence available that is derived from disparate data sets can readily be assessed using our graphical approach.

It is important to consider the available information from the point of view of both what is known and what is not known. Approaches to assessing uncertainty relevant to hazard characterization and risk assessment (reviewed by Hart et al. 2010) were originally developed for exposure and are therefore adapted to quantitative questions, whereas many of the questions posed for toxicity tend to be categorical. Types of uncertainties are most often classified into 3 categories (IOM 2013) and include uncertainties associated with variability and heterogeneity, model and parameter uncertainties, and deep uncertainty. Uncertainty that is due to inherent natural variations can be improved by making additional measurements but cannot be eliminated by further research. The evaluations of both the reliability and relevance of individual studies and the collective body of evidence offer opportunities to consider variability and heterogeneity (e.g., confidence intervals on effect response, range of doses tested, and if available, ecological heterogeneity via higher-tier studies). Model and parameter uncertainties can be reduced by further research and are therefore related to the quantity and diversity of the evidence. Deep uncertainty is present when the nature of underlying environmental processes is not understood and additional research will not resolve the uncertainty within the time frame in which a decision is made. These should be acknowledged during problem formulation and when considering coherence of the evidence.

During the integration of evidence, the graphical representation of the reliability and relevance of all included studies for different LoEs (e.g., Figure 2) allows a good visual first impression of the quantity and diversity of data available and the credibility of the body of evidence. The summaries of consistency and coherence of the evidence will also inform the confidence of individual assessors in the conclusions and recommendations. Sensitivity analyses excluding different clusters of studies (e.g., high reliability, low relevance) could be especially useful to detect how these may influence the conclusions of the WoE assessment.

DISCUSSION

In these and other applications of WoE, several barriers have hampered the widespread, consistent use of WoE, particularly when evaluating data from both ecotoxicological and environmental studies. Examples of barriers have included the lack of guidance or lack of familiarity with existing guidance, the differing goals dictated by regulatory frameworks, the additional time and resources required, and the desire for consistency and recalcitrance toward change. When focused on effects analysis, several

 DOI: 10.1002/ieam.1936

recommendations can help to overcome these and other assessment barriers.

- For the effects or hazard characterization, all data need to be evaluated in terms of reliability and relevance with a method described a priori and fit to the regulatory framework and problem formulation.
- Development and publication of a priori protocols for relevance and reliability evaluations that fit within a WoE approach tailored to the regulatory context should be a priority goal.
- The reliability and relevance of all individual studies or data should be evaluated (see companion papers on these topics, Moermond et al. (this issue) and Rudén et al. (this issue)) and presented graphically within and between the LoEs proposed here. This step becomes the basis on which to discuss and integrate the evidence available.
- The integration of the evidence should consider the credibility, quantity, and diversity of the evidence as well as its internal consistency and external coherence with the current body of knowledge within the context of remaining uncertainty (uncertainty analysis).

These recommendations will streamline the WoE process, avoiding overlap and double accounting of aspects already considered when evaluating the reliability and relevance of individual studies. The data are first integrated according to LoEs, including mechanistic insights (e.g., cellular, subcellular, genomic), in vivo experiments, and higher-tiered observational studies. Then, the data are scored or categorized on the basis of both their relevance within an LoE and their reliability. A graphical approach provides a means of illustrating visually the credibility, quantity, diversity, and consistency of specific LoEs and the body of evidence, and is helpful in deriving an expression of the confidence in the conclusions of the integration of the information. A streamlined approach to WoE for the effects or hazard characterization should facilitate reproducible and transparent assessments of data across different regulatory requirements.

Based on the aforementioned considerations, a proposed approach to incorporating reliability and relevance into ecotoxicological data evaluation as part of WoE analysis should involve the following steps. The choice of methods for each step is open to adaptation, depending on the context of the assessment and governing regulatory framework.

1) Review protection goals and develop the framework for the risk assessment (e.g., problem formulation, assessment endpoints, risk hypothesis).
2) Develop an a priori protocol for transparency and consistency. It may be desirable to develop a process whereby these protocols are published, peer-reviewed, or open to stakeholder consultation to further enhance the clarity in the approach taken.
3) Perform a well-documented literature search and gather other regulatory studies conducted for the assessment according to a systematic review process (see Whaley et al. 2016). If any preliminary screening is performed, then justification needs to be provided and documented, as well as assurance that compliance standards have been met.
4) Assess relevance and reliability of individual selected studies, as described by the companion papers of Moermond et al. (this issue) and Rudén et al. (this issue). For example, a rubric or scoring system can be established for relevance and reliability. The Supplemental Data provide an example of a rubric that could be used.
5) Integrate and evaluate the evidence according to predetermined specific LoE. For an effects characterization, the LoEs can be divided into broad generic study types such as mechanistic insights, experimental in vivo evidence for effects on typical apical endpoints and species (which can be broken into taxonomic groups and/or environmental compartments), and higher-tier studies (e.g., microcosms or mesocosms, field studies, and ecoepidemiological investigations).
6) Integrate and evaluate the evidence across LoEs to help understand the relative importance of individual LoEs to the risk hypothesis or assessment endpoints. Although inferences and expert judgment may need to be used, the end result is a characterization of the credibility of the overall body of evidence with regard to the original hypothesis.
7) Prepare an assessment of uncertainty that considers each step during the LoE assessment and identifies the influence of uncertainty on the overall conclusions. The uncertainties should be summarized and aggregated at the conclusion of the assessment.

Acknowledgment—The authors thank the Society of Environmental Toxicology and Chemistry (SETAC) for organizing the Pellston workshop "Improving the Usability of Ecotoxicology in Regulatory Decision-making." The present paper is the work of the authors and does not necessarily reflect the view or opinions of their institutions.

SUPPLEMENTAL DATA

Supplemental Data 1. An illustrative example of reliability and relevance criteria that can be used in an effects or hazard weight-of-evidence (WoE).

Supplemental Data 2. Comparison of how weight-of-evidence (WoE) criteria developed by Bradford Hill (1965) and Cormier et al. (2010) are integrated in the proposed framework.

REFERENCES

Bradford Hill A. 1965. The environment and disease: Association or causation? President's address. *Proc R Soc Med* 58(9):295–300.

Cormier SM, Suter GW, Norton SB. 2010. Causal characteristics for ecoepidemiology. *Hum Ecol Risk Assess* 16:53–73.

Hart A, Gosling JP, Boobis A, Coggon D, Craig P, Jones D. 2010. Development of a framework for evaluation and expression of uncertainties in hazard and risk assessment. York (UK): Food and Environment Research Agency. FSA Project Number T01056.

Hope B, Clarkson J. 2014. A strategy for using weight-of-evidence methods in ecological risk assessments. *Hum Ecol Risk Assess* 20(2):290–315.

[IOM] Institute of Medicine. 2013. Environmental decisions in the face of uncertainty. Washington (DC): National Academies. 280 p.

Moermond C, Beasley A, Breton R, Junghans M, Laskowski R, Solomon K, Zahner H. 2017. Assessing the reliability of ecotoxicological studies: An overview of current needs and approaches. *Integr Environ Assess Manag* 13:640–651.

Rudén C, Adams J, Ågerstrand M, Brock TCM, Buonsante V, Poulsen, V, Schlekat CE, Wheeler JR. 2017. Assessing the relevance of ecotoxicological studies for regulatory decision-making. *Integr Environ Assess Manag* 13:652–663.

Suter GW, Cormier SM. 2011. Why and how to combine evidence in environmental assessments: Weighing evidence and building cases. *Sci Total Environ* 409:1406–1417.

Whaley P, Letcher RJ, Covaci A, Alcock R. 2016. Raising the standard of systematic reviews published in *Environment International*. *Environ Int* 97:274–276.

 DOI: 10.1002/ieam.1936