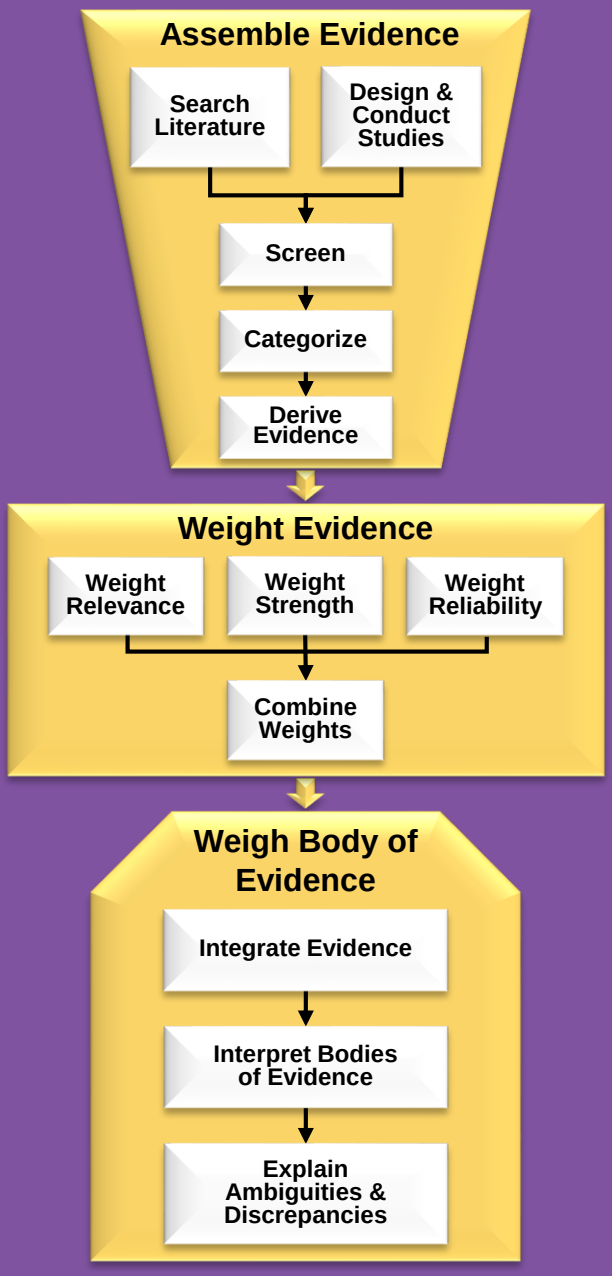
United States
Environmental Protection
Agency

EPA/100/R-16/001 December 2016
www.epa.gov/osa

Weight of Evidence in Ecological Assessment



Office of the Science Advisor
Risk Assessment Forum

EPA/100/R16/001
www.epa.gov/osa

WEIGHT OF EVIDENCE IN ECOLOGICAL ASSESSMENT

Risk Assessment Forum
U.S. Environmental Protection Agency
Washington, DC 20460

DISCLAIMER

This document has been reviewed in accordance with U.S. Environmental Protection Agency (EPA) policy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

CONTENTS

TABLES

FIGURES

BOXES

AUTHORS, CONTRIBUTORS, AND REVIEWERS

AUTHOR

Glenn W. Suter II
National Center for Environmental Assessment
Office of Research and Development
Cincinnati, OH

TECHNICAL PANEL (Contributors)

Mace G. Barron
National Health and Environmental Effects Research Laboratory
Office of Research and Development
Gulf Breeze, FL

David Charters
Office of Superfund Remediation and Technology Innovation
Office of Land and Emergency Management
Edison, NJ

Susan M. Cormier
National Center for Environmental Assessment
Office of Research and Development
Cincinnati, OH

Karen Eisenreich
Office of Pollution Prevention and Toxics
Office of Chemical Safety and Pollution Prevention
Washington, DC

Kris Garber
Office of Pesticide Programs
Office of Chemical Safety and Pollution Prevention
Washington, DC

Wade Lehmann
Office of Science and Technology
Office of Water
Washington, DC

Scott Lynn
Office of Science Coordination and Policy
Office of Chemical Safety and Pollution Prevention
Washington, DC

Chris Sarsony
Office of Air Quality Planning and Standards
Office of Air and Radiation
Research Triangle Park, NC

Glenn W. Suter II (Panel Chairman)
National Center for Environmental Assessment
Office of Research and Development
Cincinnati, OH

Risk Assessment Forum Staff

Lawrence Martin
Office of the Science Advisor
Washington, DC

EPA Peer Reviewers

Daniel A. Axelrad
National Center for Environmental Economics
Office of Policy
Washington, DC

Bruce Duncan
Office of Environmental Assessment
Region 10
Seattle, WA

Wayne Munns
National Health and Environmental Effects Research Laboratory
Office of Research and Development
Narraganset, RI

Deirdre Murphy
Office of Air Quality Planning and Standards
Office of Air and Radiation
Research Triangle Park, NC

Kathleen Raffaele
Policy Analysis and Regulatory Management Staff
Office of Land and Emergency Management
Washington, DC

Mary Reiley
Office of Science and Technology
Office of Water
Washington, DC

External Peer Reviewers

Brian W. Brooks
Center for Reservoir and Aquatic Systems Research
Baylor University
Waco, Texas

Peter M. Chapman
Chapman Environmental Strategies Ltd.
N. Vancouver, British Columbia, Canada

Valery E. Forbes
College of Biological Sciences
University of Minnesota
Saint Paul, Minnesota

Igor Linkov
Risk and Decision Science Focus Area
U.S. Army Engineer Research and Development Center
Concord, Massachusetts

PREFACE

The impetus for this document includes U.S. Environmental Protection Agency (EPA) policy, outside advice, and the expressed needs of EPA ecological assessors. Ensuring and maximizing the quality, objectivity, utility and integrity of information disseminated by the EPA "involves a 'weight-of-evidence' (WoE) approach that considers all relevant information and its quality, consistent with the level of effort and complexity of detail appropriate to a particular risk assessment" (U.S. EPA, 2002b). The EPA Science Advisory Board recommended that the Risk Assessment Forum's Ecological Oversight Committee consider the development of guidance for using the WoE approach a high priority (SAB, 2012). This document was prepared for EPA ecological assessors who expressed a desire for assistance in determining which WoE approaches are potentially appropriate for their assessments (U.S. EPA, 2010b).

This document has three goals. The first is to assist ecological assessors who plan to weigh multiple pieces of scientific evidence. We provide a standard framework consisting of three steps: assemble evidence, weight evidence, and weigh the body of evidence. We also present a broadly applicable system for assigning weights by evaluating and scoring the evidence and for combining the weighted evidence to determine the best-supported hypothesis. Additional material addresses how to deal with bodies of evidence that are discrepant or anomalous and how to express confidence in inferences based on the weight of evidence. Finally, we briefly address the weighing of evidence to derive quantities used in or generated by assessments.

The second goal is to make the weighing of evidence in ecological assessments more formal and defensible. Weighing evidence is best performed using pre-existing and well-described methods. The standard framework and default approach to weighing evidence presented in this document are expected to improve the practice and acceptance of weighing evidence.

The third goal is to make the logic of weighing evidence clearer and more consistent. For example, many of the EPA's weight-of-evidence analyses are based on Hill's considerations, a 50-year-old list that mixes characteristics of causal relationships, types of evidence, and properties of evidence. This document addresses those aspects of evidence separately and uses them systematically.

This document provides methods for weighing ecological evidence. Use of the methods will improve the consistency and reliability of WoE-based assessments and the defensibility of scientific input to decision making. This guidance is not meant to be prescriptive, nor does it dictate methods for specific programs and applications.

Tables, figures, and text boxes throughout the document provide examples of WoE practices. The specific methods in the examples were designed for particular statutory contexts and might need to be adapted before they are used in other contexts.

This document was prepared under the auspices of EPA's Risk Assessment Forum. The Risk Assessment Forum was established to promote scientific consensus on risk assessment issues and to incorporate this consensus into appropriate risk assessment guidance. To accomplish this purpose, the Forum assembles experts from throughout EPA in a formal process to study and report on these issues from an Agency-wide perspective.

ACRONYMS AND ABBREVIATIONS

BBN	Bayesian belief network
EPA	U.S. Environmental Protection Agency
FIFRA	Federal Insecticide, Fungicide, and Rodenticide Act
IRIS	Integrated Risk Information System
ISA	Integrated Science Assessment
LC_{50}	median lethal concentration
LOAEL	lowest-observed-adverse-effect level
MCDA	multi-criteria decision analysis
NE	No Evidence
NOAEL	no-observed-adverse-effect level
OECD	Organization for Economic Cooperation and Development
PRG	preliminary remedial goal
QA	quality assurance
QAPP	Quality Assurance Project Plan
r	correlation coefficient
RQ	risk quotient
SMAV	species mean acute value
SSD	species sensitivity distribution
TMDL	total maximum daily load
WoE	weight of evidence

EXECUTIVE SUMMARY

This document presents an approach for using weight of evidence (WoE) in ecological assessments. WoE integrates multiple pieces of evidence to infer a quality such as causality, teratogenicity, impairment, protection, or recovery. WoE can also be used to derive quantities such as a benchmark value, a magnitude of effect or a bioaccumulation factor when multiple estimates are available. WoE is essential for ecological risk assessment, because diverse laboratory and field information must be assembled, evaluated and integrated. The EPA has often employed WoE in ecological assessments, but, in the absence of guidelines, the methods are inconsistent and often informal. Advisory bodies, professional societies and the assessment science literature all encourage more and better WoE. The proper use of an explicit WoE methodology can mean the difference between an ecological risk characterization with a murky rationale and one that is persuasive and consistent with best practices. The system presented here to improve the practice and acceptance of weighing evidence is broadly applicable for evaluating and scoring evidence and combining the weighted evidence to determine the best supported risk characterization. By helping to make the logic of weighing evidence clearer and more consistent, WoE will help make ecological assessments more informative and defensible.

Ecological assessments that employ WoE have made numerous important contributions to environmental protection. The following five examples of high profile ecological assessments demonstrate the value added through incorporating a WoE methodology. 1)The Bristol Bay, Alaska watershed assessment to protect the world's best remaining wild salmon populations, and the people and wildlife that depend on them, used formal WoE analyses to integrate evidence from laboratory tests, field studies, and effects of similar mines to estimate risks to salmon of various mining activities. 2) The restoration of thousands of biologically impaired waters uses a WoE method to determine the cause of impairment (https://www3.epa.gov/caddis). 3) The assessment to support the definition of waters of the United States used WoE to show how different types of aquatic systems are connected. 4) The protection of ambient water quality depends on the ecological assessments that derive aquatic life criteria. The recent benchmark and proposed criteria method for major ions measured as conductivity are based on effects observed in the field. They depend on WoE to determine that the relationships in the field are not confounded and that the same value applies to different areas. 5) The high profile assessment of risks to pollinators from the neonicotinoid pesticide imidicloprid relies on weighing evidence from laboratory tests, semi-field tests, crop applications, and reports from bee keepers. WoE has been increasingly important to ecological assessment and Agency decisions because of the use of eco-epidemiological approaches.

Although WoE has been used in various types of assessments for programs across the U.S. Environmental Protection Agency (EPA), approaches have varied. Many of the EPA's WoE applications use a narrative approach with some guidance provided by lists of considerations, but that approach has been deemed inadequate by the National Research Council (NRC, 2014). The approach in this document is more formal. It provides a framework, a set of properties of evidence, a scoring system, tables for presenting results of weighting, and weighing evidence, a system for organizing evidence in terms of types and the characteristics they address and a means of dealing with ambiguous or discrepant results.

The framework and methods in this document also provide an integrated approach to both infer a quality of interest and estimate an associated quantitative value. For example, WoE can be used to weigh the evidence that a chemical causes malformations in fish in ambient exposures (qualitative) and to derive the best estimate of a benchmark concentration for the chemical (quantitative). Unlike qualitative WoE, quantitative WoE is not an established practice in the EPA. Therefore, quantitative methods are only briefly discussed.

The frameworks for qualitative and quantitative WoE are presented in Figure S-1 and Figure S-2. Although the qualitative processes might appear complex, the basic framework is simple: assemble the evidence, weight the evidence, and weigh the body of evidence. The document explains how to perform each step (Sections 4−6) and explains when and how the process can be adapted or abbreviated (Section 7). Figure S-2 previews Section 8, which presents the process for using WoE to estimate quantities.

The sections preceding the explanation of the process include an introduction that defines and explains WoE (Section 1), a description of the uses of WoE in the EPA's ecological assessments (Section 2), and an explanation of how to perform planning and problem formulation for assessments that includes WoE (Section 3). Sections following the process sections describe the complementary relationship between WoE and uncertainty (Section 9) and the path forward for implementing WoE in the EPA (Section 10). Section 11 presents the literature cited in support of the document.

Appendix A presents a glossary of terms, as used in this document, related to WoE. Additional appendices briefly review WoE methods for estimating quantities (Appendix B), selecting models (Appendix C), and inferring qualities (Appendix D). They provide context for the methods in this document. Appendix E explains that qualities such as causation have characteristics such as co-occurrence that can be used to organize bodies of evidence and to provide a basis for determining the completeness of the body of evidence. Characteristics of causation are established, but characteristics of other inferred qualities such as protective or impaired are new and have not been tested.

The result of the qualitative WoE process (Figure S-1) is a qualitative conclusion (e.g., the metal mixture is the most likely cause) and an overall weight that expresses confidence in the conclusion (e.g., weight of evidence for the conclusion is moderate) or multiple weights (e.g., relevance of the evidence is low, but strength and reliability are both high).

The result of the quantitative WoE process (Figure S-2) is a weighted quantitative value. In a WoE analysis, those results could consist of the value and its units (the proposed remedial goal is 2 mg/kg dry sediment), one or more expressions of weight (the overall weight of evidence is high) and the estimated statistical scatter of the value (the 95% confidence interval [CI] = ± 0.5 mg/kg dry sediment).

This document provides methods for weighing ecological evidence. Use of the methods will improve the consistency and reliability of WoE-based assessments and the defensibility of scientific input to decision making. This guidance is not meant to be prescriptive, nor does it dictate methods for specific programs and applications.

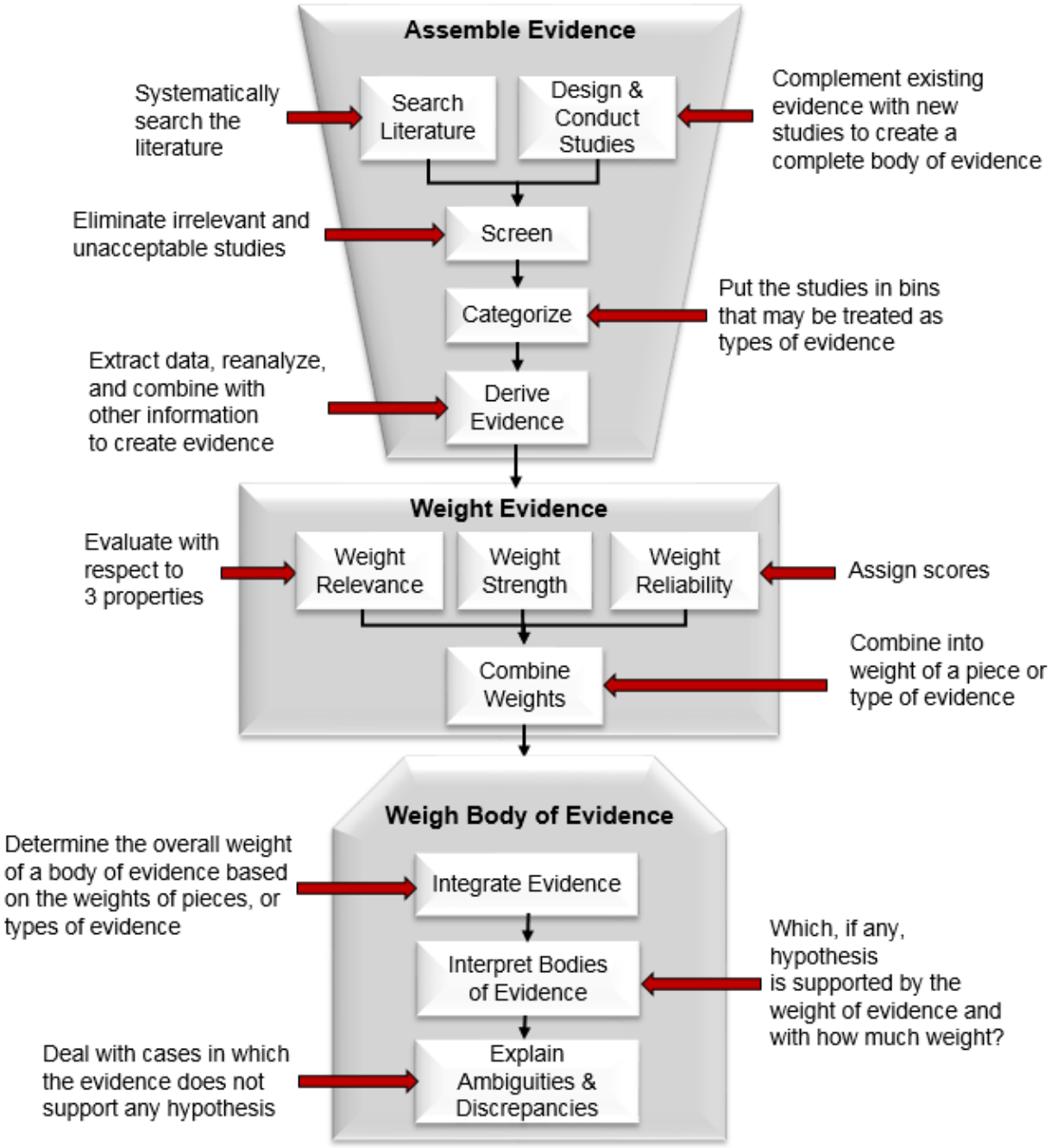


Figure S-1. An annotated diagram of the process for WoE to infer qualities by assembling evidence, weighting it, and weighing the resulting body of evidence, explained in Sections 4–7.

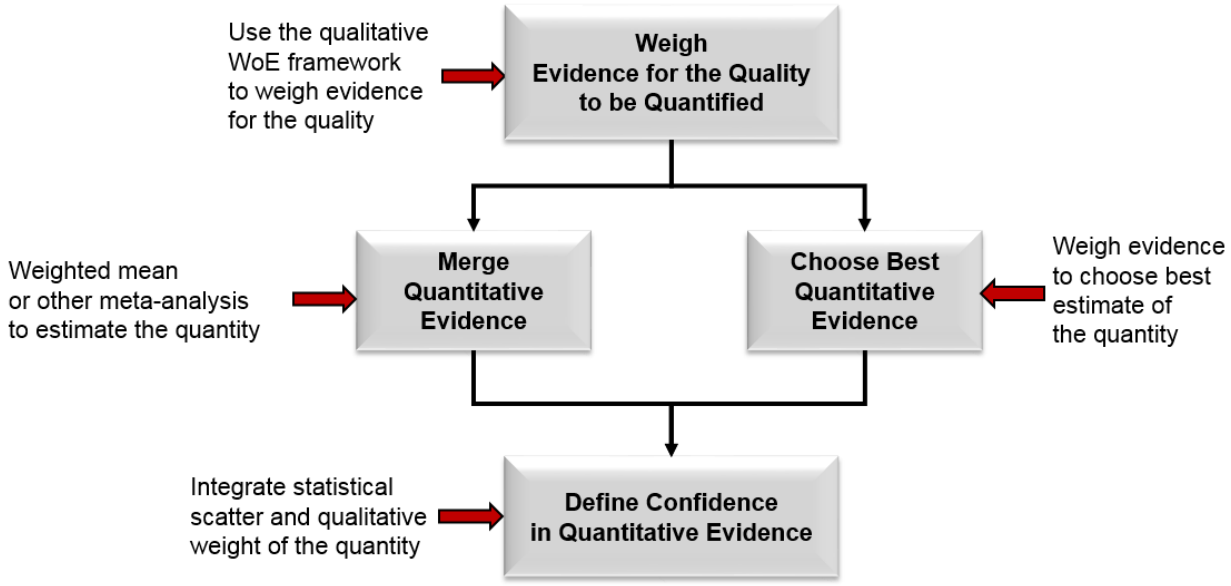


Figure S-2. An annotated diagram of the process for using WoE to estimate quantities, explained in Section 8. It begins by weighing evidence to determine the quality of interest (a qualitative WoE as depicted in Figure S-1), then uses one of two alternative methods to obtain results from multiple quantitative estimates, and finally defines the confidence in the result.

1. INTRODUCTION

The process of weighing multiple pieces of evidence is useful for judging the truth of hypotheses, identifying the best explanation of phenomena, or deriving best models or best estimates. The weight-of-evidence (WoE) process involves (1) assembling evidence, (2) weighting evidence with respect to properties, and (3) weighing the integrated body of evidence. This WoE process is embedded in larger assessment processes, which include planning, problem formulation, analysis, synthesis, and communicating results.

Scientific investigations of a topic often generate multiple pieces of evidence of various types. For that reason, WoE is inherent in medicine, engineering, and other applied sciences, including environmental assessment. Because weighing evidence is so common, the importance of the process is often overlooked and as a result, it is generally performed informally and presented as a narrative. WoE, like other scientific activities, depends on transparency for its credibility. This document encourages more explicit WoE methods, but it also encourages fitting the WoE process to the assessment to avoid processes that complicate the assessment without enhancing the results.

The *Guidelines for Ecological Risk Assessment* did not include WoE (U.S. EPA, 1998). Instead, guidance was provided for a lines-of-evidence process that is equivalent to a narrative WoE guided by considerations (see Appendix D.2). Assessment practices now have advanced sufficiently to recommend WoE for ecological assessments.

1.1. Definitions

WoE is a metaphor adapted from jurisprudence, in which multiple pieces of evidence are metaphorically placed in the pans of the scales of justice and the side with the greatest weight prevails. The metaphor is appropriate for any situation in which multiple and diverse evidence are evaluated to reach a conclusion.

In this document, WoE is defined as an inferential process that assembles, evaluates, and integrates evidence to perform a technical inference in an assessment. WoE methods have been derived to estimate a quantity (Appendix B), inform model selection (Appendix C), or reach qualitative conclusions in an assessment (Appendix D).

As part of that inferential process, WoE characterizes properties of pieces of evidence and of bodies of evidence. First, WoE determines the degree of support for a hypothesis that a piece or type of evidence provides (i.e., the weight of a piece of evidence dropped into a pan of the scales). Hence, weights indicate which pieces and types of evidence make the greatest contribution to the inference. Second, WoE determines the degree of support for a hypothesis, relative to alternatives, that the available body of evidence provides (i.e., the accumulated weight in one pan relative to the other). These cumulative weights not only inform inferences, they also indicate how much confidence assessors have in the conclusion (Section 9).

For example, "WoE was used to determine the likely cause of a bird kill" (the process). "The occurrence of carbofuran granules in the crops of dead birds is convincing evidence of exposure" (a property of a piece of evidence). "The body of evidence consistently supports carbofuran as the cause" (a property of the body of evidence). Further discussion of these terms is presented in Box 1-1. Other terms are defined in the glossary in Appendix A.

1.2. After 50 Years, Advancing Beyond Hill's Considerations

The touchstone of WoE in health and environmental assessments is Hill's considerations (Hill, 1965). Hill recognized that causal inference requires qualitative WoE because no amount of quantitative analysis of associations can suffice (i.e., correlation is not causation). He listed nine considerations to guide the process. Hill's considerations include, but do not distinguish, characteristics of causal relationships (e.g., temporality), types of evidence (e.g., experiment), properties of evidence (e.g., strength), and properties of bodies of evidence [e.g., consistency; (Cormier et al., 2010)]. EPA has adopted Hill's considerations for hazard identification in Integrated Risk Information System (IRIS) assessments and Integrated Science Assessments (ISAs).[1] Hill's considerations were designed for determining the causes of observed epidemiological effects, however, and not for applications in which effects have not been observed. In particular, Hill presumed that co-occurrence of the putative causal agent and the effect of interest (smoking and lung cancer, in his case) was clearly documented, and his goal was to demonstrate that the relationship is causal. We, however, must also address cases for which a clear real-world association has not been demonstrated.

Box 1-1. Weight versus Weigh, Weighting versus Weighing

The terms weight, weigh, weighting, and weighing, which are used throughout this document, can be confusing. This ambiguity results in part from the fact that weight is both a noun and a verb.

To *weight* a piece of evidence is to assign importance to it (a process designated by a verb). As a result of that action, the evidence has weight (a result designated by a noun). The *weighting* process formalizes the evaluation of evidence, by assigning a descriptor or score.

To *weigh* a body of evidence is to combine the weights assigned to each of the pieces (a process designated by a verb). As a result of *weighing*, the body of evidence has weight (a result designated by a noun). The weighing process formalizes the integration of evidence by determining the weight of the body of evidence.

The WoE system presented here treats explanatory implication of evidence, types of evidence, properties of evidence, and properties of bodies of evidence as separate aspects of WoE (Box 1-2). Because each serves a different function in the WoE process, they are not treated as equivalent considerations. Distinguishing them makes clear that Hill's list lacks essential characteristics, types, and properties of evidence. These aspects are the basis for weighting and weighing evidence for causation and other qualities as presented in Sections 5–7.

[1] IRIS is a program that evaluates information on health effects of environmental contaminants to determine what hazards they pose to humans and to develop benchmark values (http://www.epa.gov/iris/). ISAs assess the scientific evidence for health and welfare effects of criteria air pollutants (http://www.epa.gov/ncea/isa/).

Box 1-2. Aspects of Evidence

This approach to weighing evidence distinguishes aspects of evidence in the WoE process that may not be clearly distinguished in prior approaches. They are introduced here and explained in more detail as they appear in the WoE process.

Piece of evidence: A piece is the evidence derived from a particular experiment or observational study. A piece of evidence is the minimum unit that might be weighted.

Type of evidence: A type is a category of evidence based on the nature of the study from which the evidence is derived. Examples include single-species, laboratory toxicity tests; microcosms; mesocosms; biomarker surveys; and community surveys. Types are used to organize and combine pieces of evidence to simplify weighting and weighing.

Explanatory implication of evidence: An explanation expresses the logical relationships between evidence and inference. For example, laboratory toxicity tests may provide evidence of the sufficiency of an exposure to cause the effect, and community surveys may provide evidence of co-occurrence of the effect and putative cause.

Body of evidence: A body of evidence is all of the evidence that applies to a particular hypothesis.

Property of evidence: Properties are the aspects of evidence that determine how much weight (influence) it should have. This approach uses three general properties: relevance, strength, and reliability.

Property of bodies of evidence: Bodies of evidence are weighted with respect to collective properties including number, coherence, diversity, and absence of bias.

1.3. SCOPE

This document is intended to inform ecological assessments. Human health assessments are mentioned and cited only as background and for purposes of comparison. Weighing multiple types of evidence has been performed more widely in ecological assessments than in human health assessments (Krimsky, 2005; Suter, 1993). Although ecological and human health assessments can be performed similarly, the types of ecological assessments that have driven the development of WoE tend to differ in important respects from human health assessments. In particular, types of ecological studies that contribute evidence to Superfund and Clean Water Act assessments include effluent and ambient media toxicity tests, toxicity identification evaluation, in situ tests, biotic community surveys, and demographic models. In addition, ecological assessments must weigh evidence related to multiple endpoint species and levels of organization, which increases the need for formal WoE methods.

This document presents a framework for application of WoE in various assessment contexts. WoE is an assessment tool, just as modeling and statistical analysis are assessment tools. Therefore, the framework for WoE is not a substitute for assessment frameworks (see Section 3).

WoE can be carried through the entire assessment process or be applied to an inference that is only a component of the assessment. WoE can be limited to hazard identification (e.g., Does ozone reduce plant growth at ambient levels?), the determination of an assumption (e.g., Should a bioaccumulation factor be used?), or the estimation of a parameter (e.g., weighing field and laboratory estimates of a chemical's half-life in water). In contrast, a WoE process can carry through an entire condition or causal assessment.

This document cites literature that is directly relevant to explaining this approach to WoE but does not attempt to cover the history of WoE in environmental assessments. Reviews of WoE approaches are available elsewhere (Rhomberg et al., 2013; Linkov et al., 2009; Pope et al., 2007; Krimsky, 2005; Weed, 2005).

This document focuses on WoE for qualities such as causality and impairment (Box 1-3). WoE methods to derive quantities and to choose models are discussed in Appendix B and Appendix C, and the use of qualitative WoE to enhance the derivation of quantitative results is discussed in Section 8. Qualitative WoE is emphasized because it is more common in ecological assessments than is quantitative WoE (Appendix D).

Box 1-3. Qualities and Qualitative Weight of Evidence

Qualitative WoE: A qualitative WoE is an assessment for which the endpoint is a quality such as a condition, mode of action, source, or type of effect. It typically includes both quantitative and qualitative evidence, but the result is not a numerical quantity.

Quality of Interest: The quality for which evidence is weighed is most often causality, but it can be any quality of interest. Examples of qualities that can be or have been inferred by WoE include:

- Coxie Creek is impaired.
- Acid mine drainage causes the impairment.
- Polychlorinated biphenyls are bioaccumulative.
- Selenium is teratogenic in fish.

Quality of Evidence: This document does not recommend weighting the quality of evidence. Instead, specific properties and sub-properties of evidence are weighted. This is because the term quality has been found to be too broad to be useful in the weighing of evidence (Higgins and Green, 2011). However, quality is used in a broad sense that is consistent with U.S. government data quality policy (U.S. EPA, 2002b).

1.4. Benefits and Challenges of Weight of Evidence

Although this document recommends the use of WoE, it recognizes that WoE can have limitations in practice. We assume that assessors will consider the potential benefits and challenges of WoE (Box 1-4) relative to the requirements of their assessment before proceeding to implement the approach recommended here.

WoE techniques inevitably involve subjective expert judgments. Such judgments can cause WoE to be criticized for being biased or arbitrary. However, the objections to subjectivity can be diminished by some good practices.

1. Prior to the assessment, specify the WoE method in as much detail as is practical to minimize the need for improvised judgments about methods or assumptions during the assessment.

2. Use the standard judgments of a program expressed as standard criteria for assigning or integrating weights. For example, tests performed using an Organization for Economic Cooperation and Development (OECD) protocol with good laboratory practices might be given a standard score of +++ for reliability.

3. Be objective, in the sense of being unbiased, by self-auditing (Box 1-5).

4. Work in groups and attempt to achieve consensus concerning judgments.

5. When making judgments, try to represent the opinions that knowledgeable and unbiased members of the scientific community would have, given the evidence and the inference to be made.

6. Find assessors with sufficient relevant knowledge and experience to qualify as experts.

Box 1-4 Potential Benefits and Challenges

Inference by weighing multiple pieces of evidence has advocates who recognize its potential benefits, but challenging aspects of its application have led some assessors to oppose its use.

Potential Benefits of WoE

The primary potential benefit of WoE is the greater confidence in results obtained by considering all relevant and reliable evidence. For example, it is not uncommon for causal assessments to consider only statistical evidence of co-occurrence of an effect and its potential causes. This approach provides much less confidence than one that also considers evidence of temporal sequence, interaction, and other characteristics of causal relationships. In many cases, no single type of evidence is sufficient to reach a conclusion. Ecological assessments of polluted ecosystems commonly benefit from considering evidence from laboratory toxicity tests of chemicals, tests of effluents, or ambient media and biological surveys. This benefit of WoE occurs because the body of relevant and scientifically credible evidence provides a more complete picture than does any piece of evidence alone.

A second potential benefit is an increase in the defensibility of an assessment. An explicit WoE method demonstrates that all relevant evidence has been considered and no credible evidence, either in support of or contrary to a hypothesis, has been arbitrarily dismissed. Without an explicit process planned in advance, reviewers might criticize or even dismiss an assessment for excluding data or evidence that they believe should have been given more consideration.

Defensibility of assessments also might be increased by the transparency of the processes it uses for inference. A formal WoE method enables reviewers and readers to understand and critique the processes of assembling, weighting, and weighing the evidence.

Potential Challenges to WoE

A formal WoE process can require considerable time and effort, which could lead to performance of fewer assessments or delayed decisions. Completing and documenting a formal systematic literature review or implementing an evidence scoring system might not be resource effective if the same conclusion can be obtained with a less resource-intensive assessment. The solution to this challenge is to tailor the WoE method to the assessment.

Box 1-5. Subjectivity and Objectivity

WoE-based assessments rely on subjective professional judgments because there is no other means of weighing a diverse body of evidence from models, laboratory tests, field tests, field surveys, and other information sources to identify the best-supported hypothesis. Although inevitable, subjectivity often is considered undesirable in assessments as in other scientific contexts.

Objective properties, such as the number of fish species in a stream, are properties external to the investigator that can be confirmed by any other investigator. Subjective properties such as the reliability of biological survey results are opinions of the investigator evaluating the survey, not inherent properties. The extent to which assessors agree in their judgment of reliability is due to shared preferences, not a measurable attribute of the surveyed community.

Subjective inferences can be objective in another sense, which is defined in federal policy (U.S. EPA, 2002b): Any inference can be considered objective if it is performed in a disinterested—and therefore impartial—manner. The guidance in this document aims to create circumstances for which an inference is minimally influenced by any personal or institutional biases and in which any systematic bias can be detected.

Performing laboratory or field studies to generate data for multiple types of evidence (e.g., chemistry, toxicity, and biology) for a WoE analysis—as opposed to simply summarizing data from available literature—is more resource intensive than generating a single type of evidence (e.g., only chemistry). The value of new information to the assessment should be considered during the planning stage to obtain sufficient, but not surplus, evidence (Keilser et al., 2014).

Complex WoE methods can obscure rather than clarify the derivation of results, particularly if the method is not clearly presented. A reader is likely to dismiss the results if the assessment is incomprehensible. Clear and consistent methods can reduce this problem.

Finally, a fundamental objection to WoE is that it might mix scientifically less robust evidence with highly reliable evidence. In some cases, the body of at least minimally relevant and reliable evidence might not contribute information beyond that supplied by a single best piece of evidence. This issue is addressed by including explicit steps for screening and weighting evidence rather than treating all evidence as equal.

These objections can be minimized by careful planning and by clearly presenting the WoE process. The WoE methods should be appropriate to the assessment. The amount of detail in evaluating and scoring evidence should be appropriate to the amount and diversity of the evidence, the time and resources available, and the degree to which decision makers and stakeholders wish to engage in a WoE process. The amount of detail can also be based on the degree to which the results of an assessment are potentially contentious. If, for example, an informal review of the evidence clearly shows a site is highly toxic and the obvious decision will be to remediate, a more detailed WoE process might waste resources, prolong the contamination, and confuse the issue. In situations for which the decision is obvious and could be urgent, a protracted WoE process is counterproductive.

2. APPLICATIONS OF WEIGHT OF EVIDENCE

The uses of WoE in ecological assessments are diverse, in part because ecological assessments are conducted in diverse regulatory contexts that require different assessment methods. In addition to asking assessors for predictions of potential adverse outcomes of proposed actions (i.e., to perform risk assessments), decision makers ask assessors to determine conditions (i.e., to perform condition assessments), determine the likely causes of adverse conditions (i.e., to perform causal assessments), and determine the actual outcomes of actions [i.e., to perform outcome assessments (U.S. EPA, 2010b; Cormier and Suter, 2008)]. Each type of assessment has its own logic and methods, and therefore, weighs evidence somewhat differently. Figure 2-1 illustrates a way to organize such assessments. Assessments can be initiated by the results of a prior assessment, and WoE results from prior assessments can inform subsequent assessments. For example, if a condition assessment identifies a biological impairment, a causal assessment should follow to identify the cause and source of the causal agent, a predictive risk assessment should follow to determine the remedy, and an outcome assessment should determine whether the remedy has sufficiently improved conditions. Assessments also can be initiated by external demands. For example, a predictive risk assessment might be prompted by an application to market a new pesticide.

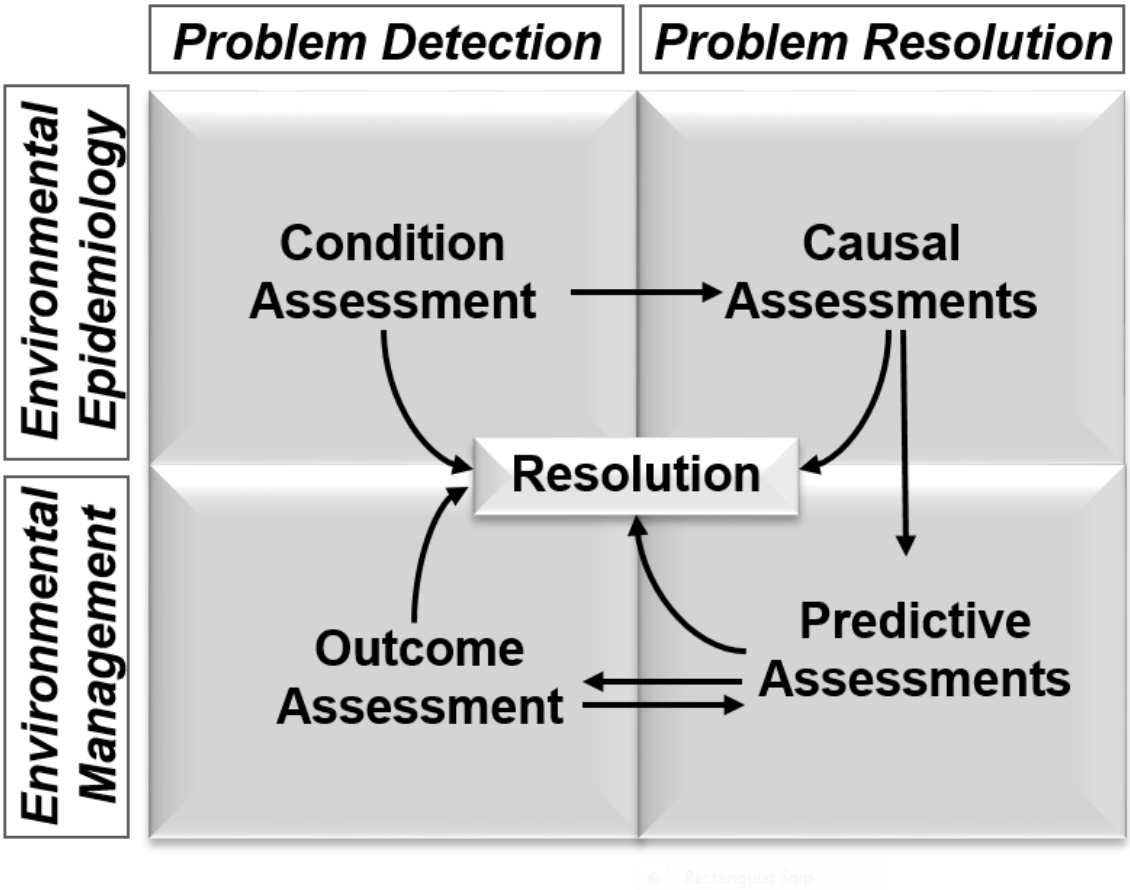


Figure 2-1. A framework depicting the relationships among types of environmental assessments. Modified from U.S. EPA (2010b); Cormier and Suter (2008).

All types of ecological assessments can weigh evidence quantitatively to combine multiple numerical values or to choose a model (see Appendix B and 11.Appendix C). This document, however, emphasizes using WoE for integrating evidence to derive a qualitative conclusion that supports a decision. Examples of qualitative conclusions are the biotic community of a stream is impaired, selenium causes cranial deformities, low pH caused the impairment, or coal ash is the source of the arsenic. To determine what qualities might require WoE, it can be helpful to consider what qualitative questions are answered by the various types of environmental assessments (Box 2-1).

Environmental problem solving often requires sequences of inferences to derive qualitative results and then quantitative results. For example, an ISA, that qualitatively weighs evidence to determine whether relevant concentrations of the pollutant cause particular effects, precedes the development of National Ambient Air Quality Standards. A qualitative weighing of evidence to determine whether the contaminated medium is causing significant adverse impacts precedes development of quantitative cleanup goals for a contaminated site. A qualitative weighing of evidence to determine whether a water pollutant is causing biological impairment might precede the development of quantitative total maximum daily load (TMDL) for the pollutant. These qualitative assessments could be treated as a separate product from the quantitative assessment (e.g., a qualitative ISA before a quantitative Welfare, Risk and Exposure Assessment). Alternatively, the qualitative assessment might be nested within the quantitative assessment as part of the problem formulation (e.g., as a hazard identification).

This section describes the various types of assessments, their application by EPA and the current and potential roles of WoE in performing the assessments.

2.1. Contaminated Sites

Contaminated sites provide a wide scope for applying WoE because the contaminated and potentially biologically impaired conditions allow for the generation of diverse evidence from observation, sampling, analysis, and testing. As a result, most of the relevant methods and literature on ecological WoE address contaminated sediments, soils, and waters (Appendix D). In the EPA, the primary venue for contaminated site assessments is Superfund sites.

The goal of contaminated site assessments is to determine whether contaminants pose an unacceptable ecological risk and, if so, what should be done to reduce it. When assessing a site, specific impaired areas can be identified without WoE by comparing contaminant concentrations to screening benchmarks. Those benchmark concentrations also could serve in some cases as the remedial goals, allowing for completion of the assessment without weighing evidence. For Superfund sites, however, multiple types of site-specific evidence are typically collected, including results of toxicity tests of contaminated media, biological surveys, or analysis of biological samples for biomarkers or body burdens of contaminants (Luftig, 1999). Such bodies of evidence can be weighed in a merged condition and causal assessment to answer the question: Is an unacceptable risk associated with the site contaminants? Such a qualitative assessment is the topic of the approaches for the sediment quality triad and the Massachusetts and Oak Ridge National Laboratory WoE methods described in Appendices D.3, 11.D.5, and D.7. Such assessments can also be performed using the approach presented in the following sections.

Box 2-1. Qualitative Questions for Which Evidence is Weighed in Different Types of Assessments

Causal Assessment

Causal assessments identify causes of observed effects and they take one of two general forms.

What causes general effect y? This is the general causal question. Examples: What causes colony collapse disorder in honeybees? What causes coral bleaching?

What causes specific effect y? This is the specific causal question. Examples: What caused low invertebrate species richness in the Coal River? What caused the 1980s decline in San Joaquin kit foxes in Elk Hills?

Risk (Predictive) Assessment

Most EPA predictive assessments are risk assessments. They begin with a problem formulation that includes a causal question regarding what hazard is associated with an agent and could be assessed.

Does agent x cause general effect y? This is a general hazard identification question for assessments like setting benchmarks or permitting use of a chemical. Examples: Does smoking cause lung cancer? Does atrazine cause deformities in frogs?

Might agent x cause effect y? This is a hazard identification question asked in risk assessments for specific actions. Examples: Might a tailings spill at the proposed mine cause a loss of salmon spawning? Might permitting a new use of atrazine result in an increase in frequency of frog deformities?

Is agent x causing effect y? This is a hazard identification question asked in risk assessments for an agent under existing conditions. Examples: Are the sediment contaminants reducing invertebrate species richness or abundance? Is acid deposition reducing forest production?

These causal questions can be answered by a separate causal assessment prior to the risk assessment or in the problem formulation for the risk assessment. Without hazard identification, defining the endpoint entity and attribute has no basis beyond using default endpoints such as fish mortality.

Other qualitative questions that might involve WoE also are answered in risk assessments. For example:

- Should dietary exposure be considered?
- Will the endangered species occur on the site in the future?
- Is the site large enough to support an assessment population of the species of concern?
- Is long-range transport a significant source?
- What are the contaminants of concern?
- Is the chemical a persistent organic pollutant?

Condition Assessment

Qualitative questions concerning conditions, such as the following, are relatively straightforward to assess when standard definitions of conditions are provided.

- Is the ecosystem impaired?
- Is the species endangered?
- Is the water contaminated?

However, framing the question could be conceptually difficult, particularly with respect to the standard of comparison: natural conditions, reference conditions, historical conditions, or defined standards.

Outcome Assessment

Qualitative questions of concerning outcome are like condition questions but include issues regarding the efficacy of remedial or regulatory actions.

- Is bioremediation sufficiently reducing exposure of endpoint species?
- Is the liner leaking?
- Is the site acceptable for recreational fishing?

Many Superfund assessments treat pieces of evidence as independent lines of evidence without weighing them (Integral Consulting Inc., 2013). Causal assessments performed by WoE for contaminated sites are illustrated by the Elk Hills and California Gulch cases (U.S. EPA, 2011c, 2009a). Such WoE assessments can be sufficient to complete the investigation, provided the decision logic is as follows: If the sediment or soil presents unacceptable risk due to waste contamination, remediate it (e.g., cap, remove, or biotreat it). For Superfund sites, however, a separate step defines chemical-specific preliminary remedial goals (PRGs). Either site-specific values or standard benchmark values might be used as PRGs for the areas identified as impaired by WoE. Site-specific PRGs can be derived from tests of site media or field exposure-response relationships used in the qualitative WoE to identify waste-impaired areas. However they are derived, PRGs represent acceptable risks, and are typically used in selecting a final remedy.

2.2. Environmental Condition

Although the condition of all ecosystems is of concern, the Clean Water Act is unique among U.S. antipollution laws in requiring that all sites (i.e., all state waters) be assessed to determine if they are impaired. The result is each state's 303(d) lists of impaired waters. Most listings are for exceedance of water quality standards, including concentrations of chemicals, levels of biological pollutants (e.g., fecal coliform counts), and physical properties (e.g., temperature). In addition, ecological properties can serve to identify impaired waters based on biological or narrative criteria. Commonly, multiple types of biological data, termed metrics, are combined into an index (Blocksom and Johnson, 2009; U.S. EPA, 1996). The regulations indicate that states should evaluate "all existing and readily available information" in developing their 303(d) lists (40 CFR §130.7(b)(5)), which suggests a WoE approach or other approaches (such as independent applicability) that consider all information (Appendix D, Box D-1).

If a water body is declared impaired based on biological effects, the cause should be identified. Guidance for using WoE to determine the cause of biological impairment was developed for the Office of Water (U.S. EPA, 2000), and a web-based support system was developed to aid in its application (http://www.epa.gov/caddis/). This system is the principal model for the WoE approach presented in this document.

Once the cause of water body impairment has been determined, the sources can be identified so that TMDLs can be developed (Clean Water Act §§ 130.2(f)-(i) and 130.7(c)). This is similar to the practice of determining sources at contaminated sites is known as environmental forensics because it often involves establishing legal liability (Murphy and Morrison, 2002). Its methods include environmental fingerprinting, isotope analyses, tracer studies, and transport modeling, and it often uses WoE. TMDLs are developed and implemented to eliminate the impairment by apportioning pollutant loading among the sources. This step typically does not involve WoE. The TMDL process often includes evaluation of the results, which can lead to removal of the stream reach from the 303(d) list. This outcome assessment can rely on biological endpoints and could involve WoE.

2.3. Existing Pesticides and Industrial Chemicals

The reregistration of pesticides provides an opportunity to collect multiple pieces and types of evidence related to pesticide use, as well as conventional laboratory data, that is generated from the registrant's data package and literature reviews by EPA. This body of evidence, which is weighed in a WoE narrative, can include results from laboratory studies of chemistry and toxicity, mesocosms, field experiments, surveys, and incident reports; pesticide monitoring and use data; and transport, fate, and exposure modeling.

Risks to animals and plants from pesticides are assessed using a tiered approach. In general, most assessments are focused on Tier 1, intended to estimate conservative environmental exposures. The risks associated with those exposures are assessed using risk quotients (RQs), which are predicted exposure levels divided by effect benchmark levels. RQs are based on the most sensitive endpoints available for survival, growth, or reproduction. When RQs exceed levels of concern, risks are characterized using a narrative WoE analysis. Although the process for deriving RQs is well established, the WoE analysis depends on the nature of the available data and the assessed chemical and its use patterns. Examples of evidence analyzed to evaluate risk conclusions include comparing estimates of exposure from models or monitoring data to available toxicity endpoints, evaluating whether laboratory fate studies are consistent with field dissipation studies, and using species sensitivity distributions to evaluate impacts of conservative assumptions on risk conclusions.

Assessments of existing chemicals under the Toxic Substances Control Act (TSCA) can weigh diverse bodies of evidence to determine whether a chemical substance presents or could present unreasonable risk of injury to health or the environment. WoE is used when conducting risk evaluations under TSCA Section 6(b) and making risk management decisions for existing chemical substances. Existing chemical assessments typically rely on a combination of predictive modeling and other computational approaches, empirical laboratory or field data specific for the chemical being assessed and empirical laboratory or field data from analogous chemicals. For example, the body of evidence can include empirical or modeled data on use and release and on transport and fate, physical/chemical properties of the chemical in question, exposure data that include field monitoring and modeled data and hazard data from laboratory toxicity tests or modeling. This information is used to assess risk and, much like the registration of pesticides, RQs are developed and the analysis depends on the available data. The assessments rely on a set of seven study quality and selection considerations (U.S. EPA, 2015b). Implementation of the new (signed June 22, 2016) Frank R. Lautenberg Chemical Safety for the 21st Century Act may change these practices. The act expressly uses the term "the weight of the scientific evidence."

2.4. New Pesticides and Industrial Chemicals

Registration of new pesticides is based on a large and diverse body of evidence (relative to other new chemicals) pursuant to the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) pesticide registration data requirements (http://www2.epa.gov/pesticide-registration/data-requirements). Similar to risk assessments for existing chemicals, a tiered approach is used. The major difference between risk assessments for new versus existing pesticides is generally the amount and types of available data. For new chemicals, the data required under FIFRA generally are available. Incident reports, monitoring data, and toxicity data for nonstandard test species, however, are rarely available for new chemicals.

The available evidence is also weighed when reviewing, assessing, and regulating new industrial chemicals under TSCA before they enter commerce. Assessments of new chemicals rely on a combination of predictive modeling and other computational approaches and empirical data from analogous substances and chemical categories and from empirical data the manufacturer submits. As with existing chemicals, the implementation of the Lautenberg act is likely to increase the importance of WoE for new chemicals.

Review of chemicals for potential endocrine-disruptive mechanisms of action is required for all pesticide active ingredients, all food-use pesticide inert ingredients, and some industrial chemicals. A WoE approach based on a set of considerations has been used for that purpose (U.S. EPA, 2012b).

2.5. Benchmark Derivation

The derivation of numerical criteria, standards, remedial goals, screening levels, and other benchmark concentrations or doses involves performing a type of risk assessment. Qualitative WoE can be applied during problem formulation to determine what hazards should be considered in the assessment. For example, ISAs—which identify hazards that may be assessed in Welfare Risk and Exposure Assessments for National Ambient Air Quality Standards—use WoE to determine what effects on welfare occur in the relevant range of air pollutant levels (U.S. EPA, 2014b, 2013). If different methods or data sets are used to derive benchmarks for different exposure routes or modes of action, WoE can be used to determine which approach to apply, as in the development of IRIS values (human health benchmarks) for carcinogens versus noncarcinogens (U.S. EPA, 2005b). Finally, if a benchmark is derived using field data, WoE can be used to determine whether the evidence suggests the relationship is causal or confounded (U.S. EPA, 2011b).

2.6. Proposed Discharges

WoE is seldom an option when assessing risks from proposed new discharges because evidence is limited to predicted release rates of constituent chemicals used in transport models, and the resulting estimated concentrations are related to benchmark concentrations. Tests of synthetic effluents or analogies to existing effluents that are expected to be similar to the proposed discharge, however, could create bodies of evidence that could be weighed.

2.7. Special Purpose Assessments

Some assessments are performed for particular nonroutine decisions. For example, the EPA performed a watershed risk assessment to inform potential decisions concerning the mining of metals in the Bristol Bay, Alaska, watershed (U.S. EPA, 2014a). The evidence used was diverse and included analogies to prior metal mines and pipelines. The evidence was therefore weighted and weighed using a method like the one in this document. Similarly, EPA borrowed from the WoE method developed for causal assessment of impaired waters to assess the connectivity of streams and wetlands to support the definition of waters of the United States (U.S. EPA, 2015a). These examples show how one-of-a-kind assessments can adapt the general approach presented in this document.

3. PROCESSES FOR WEIGHING EVIDENCE

3.1. An Introduction to the Weight of Evidence Process

The weighing of evidence is a tool that can be used for any type of assessment, at any point in an assessment process or throughout the entire assessment (Section 2; Figure 3-1). In a risk assessment, WoE is used in the problem formulation phase to identify the hazardous properties of the agent being assessed and the appropriate assessment endpoints. In particular, evidence can be weighed to determine whether it supports the hypothesis that a chemical causes some particular effect, such as failure to hatch, effects on a particular taxon, or effects by a particular route of exposure or mode of action (Meek et al., 2014). In the analysis phase, WoE can be used to select assumptions, estimate parameters, and develop models (Section 8, Appendix B, and Appendix C). In the synthesis phase, the models and parameter estimates are used to characterize risks, determine most likely causes, or determine whether an ecosystem is impaired. When multiple exposure-response relationships, spatial/temporal associations, or other relationships have been developed, WoE is used to combine them or to determine which relationship is the best (weightiest). Other types of assessments use WoE at different points in their frameworks (Section 2). In a causal assessment, WoE is carried through all analysis and synthesis phases (Norton et al., 2014).

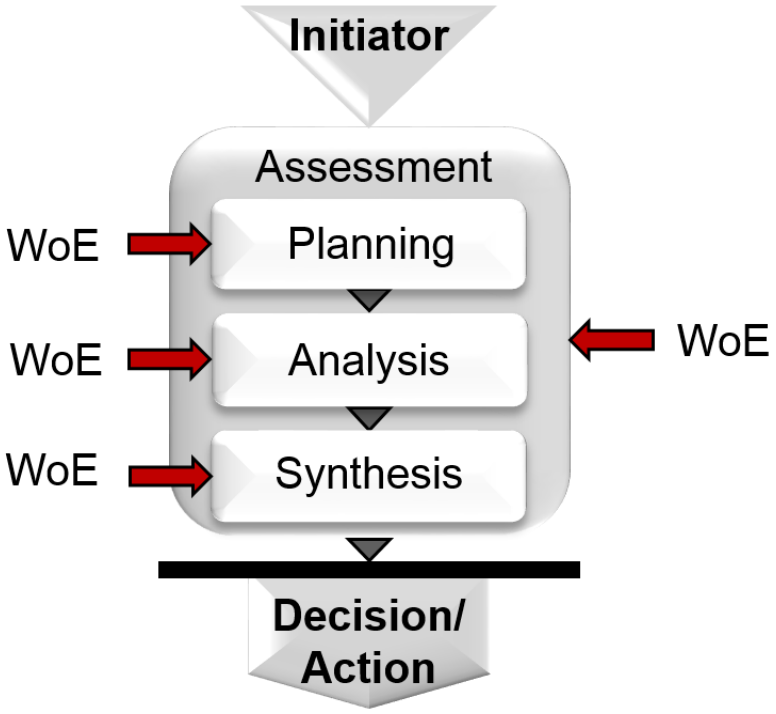


Figure 3-1. A basic framework for all types of environmental assessments (U.S. EPA, 2010b; Cormier and Suter, 2008). Weight of evidence can contribute to one or more individual steps (WoE on the left) or can be the basis for the entire assessment (WoE on the right).

However used, a formal WoE involves the same basic process (Figure 3-2). First, the evidence is assembled. This step involves finding published studies or performing new studies, determining whether their results are acceptably relevant and reliable, and extracting and analyzing data to generate useful evidence. Second, the pieces of evidence are weighted (i.e., evaluated and scored) to determine their influence on the results. Third, the body of evidence is weighed (i.e., weights are integrated and the body of evidence is interpreted) to arrive at a result.

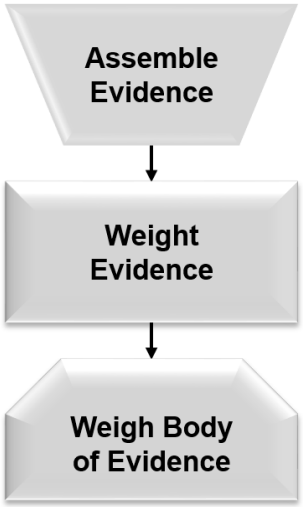


Figure 3-2. The basic WoE process (Suter and Cormier, 2011). The steps are elaborated in Sections 4–6, respectively, in this document.

WoE processes vary among applications, but these three steps are fundamental. First, you must have evidence. The final step, combining and interpreting (i.e., weighing) the evidence, is necessary to make the body of evidence inform the inference. The middle step, weighting, is also essential in that different pieces and types of evidence seldom have the same influence in an inference. However, the explicit assignment of weights is often skipped. This should be done only if the evidence is found to all be equally relevant and reliable (Section 7).

In many cases, multiple pieces of evidence are not explicitly weighed. Instead, evidence is informally weighed in a narrative (Box 3-1).

WoE can be applied to derive qualitative or quantitative results. It might be used to assess qualities such as causality, impairment, recovery, or occurrence of a specific effect or used to estimate quantitative results such as a benchmark value, a model parameter value, or the magnitude of effects. Inferences to quantities begin with quantitative evidence and apply quantitative methods (Section 8 and Appendix B). Inferences to qualities, however, use all relevant evidence (qualitative and quantitative) but apply qualitative inferential methods (Sections 5–7 and Appendix D). Even in WoE methods such as the Massachusetts system that use numerical scores, those scores are used to document qualitative judgments to reach a qualitative result (Menzie et al., 1996). The same qualitative methods can be used to make qualitative judgments about quantitative results (Section 8). For example, they might be used to judge the reliability of a quantitative effect estimate (e.g., 26 km of stream would be impaired and that result is highly reliable).

The general approach introduced in this section is intended to be flexible and broadly applicable. It is intended to provide more rigor and transparency than narrative WoE but less complexity and more flexibility than numerical systems and indices (Appendix D). The approach also is intended to merge inferences to qualitative and quantitative results in a useful manner. It can be applied to all types of environmental assessments and to both data-rich and data-poor cases.

Box 3-1. Best Practices for Narrative Weight of Evidence

Although this document explains how to use a formal method for WoE with explicit weighting and weighing, we recognize not everyone will follow this approach. Some assessors will continue to use a traditional narrative approach. In such cases, WoE narratives can benefit from some good practices.

1. When using evidence from the literature, design and carry out a literature search that will find the information needed in an unbiased manner (Section 4). Define not only the search terms but also the criteria for screening the results to obtain relevant and reliable evidence.
2. Even in a narrative approach, provide a table summarizing the selected evidence and indicate what was excluded and why.
3. Avoid narrative reviews that describes one piece of evidence after another and then present a conclusion. Instead, logically structure the narrative. At a minimum, present the evidence for a hypothesis and the evidence against it, and explain why one side has greater weight.
4. Make the narrative organization clear. Use lists and subsection headings to help the reader understand the logical structure.
5. Present the results clearly. Which hypothesis is best supported by the WoE, and how much confidence can be placed in the conclusion?
6. Express the degree of confidence in the conclusion.

3.2. Planning the Assessment to Use Weight of Evidence

All environmental assessments begin with a planning phase that defines how the assessment will be conducted. In risk assessments, this phase includes two steps termed planning and problem formulation (U.S. EPA, 1998). Assessment plans identify the evidence needs, methods that will be used to generate the evidence and methods for weighting and weighing multiple pieces and types of evidence. Specifying the WoE method in advance enhances transparency and defensibility. In particular, a previously defined system for evaluating evidence and assigning scores prevents the adjustment of weights to achieve a desired result. To the extent feasible and helpful, standard frameworks and methods should be used (Box 3-2).

In practice, there are levels of standardization. Sections 4–7 present a general WoE approach for EPA, which could reduce the burden of planning an assessment. Additional standardization can be achieved at the program level or in individual regions. The greatest specificity, however, occurs in the planning of individual assessments. Not specifying in advance how evidence will be

Box 3-2. Standardization of Weight of Evidence

Standardization of assessment practices is desirable in general. Standardization can increase efficiency by reducing the number of decisions to be made, improve assessments by setting standards of practice, and reduce bias by reducing the opportunities for assessors to make personal judgments. If the standard practices are too rigid and prescriptive given the variability in assessment problems and conditions, they can reduce efficiency, force less than optimal assessment practices, and create frustration. This conundrum is compounded when WoE approaches are used. Decisions are made concerning, among others, which inferences should be based on WoE, what evidence should be included, what properties of the evidence should be considered, and how the weights should be expressed. These decisions should be made on a program-specific basis.

weighed delays the decisions until the data are in hand and the evidence is being generated, weighted and weighed. At that point, the consequences of decisions might be apparent to the assessor. For example, if sediment toxicity tests for an area result in an average 33-percent mortality and no standard for scoring has been set, an assessor can score the evidence as 0, + or ++ (ambiguous, weak, or strong). That judgment could be biased if the assessor knows that the choice can tip the balance for or against dredging the sediment. Removing the temptation to choose a weight that gives a preferred answer by standardizing weights whenever appropriate is the better option.

Exceptions do exist. A detailed plan, on occasion, can result in nonsensical results when implemented, due to peculiarities of the case (Johnston et al., 2002). Therefore, allowing for deviations, and documenting the rationale, as Johnston et al. (2002) did, is essential.

Finally, when a decision is made to use WoE but the method was not prescribed in an analysis plan, assessors should document how they weighed the evidence and why they made their choices. Expert judgment, although essential, should be transparent.

Assessment planning should be adapted to include WoE considerations. Stakeholders and decision makers can be consulted for assurance that the weighing process is acceptable. WoE methods that are compatible with the potentially available data and evidence should be chosen. Alternatively, data needs should be specified with WoE in mind. In particular, literature reviews and new studies should be coordinated to generate complementary evidence. If a field survey will sample benthic invertebrates, literature reviews should seek toxicity studies of benthic invertebrates. Any new laboratory tests should address benthic invertebrates and preferably taxa that are sensitive or important at the site. Field sampling for different pieces of evidence (e.g., chemical concentrations, habitat characteristics, and occurrence of biological taxa) should be collocated in space and time so that the results can be used to derive relevant relationships or make comparisons.

When possible, the assessment problem to be addressed by WoE should be formulated in terms of alternative hypotheses. Identifying the hypothesis best supported by the WoE is conceptually simpler and more convincing than determining whether a single hypothesis has sufficient WoE. One hypothesis, considered alone, may appear to have sufficient evidence, but another hypothesis may have more and weightier evidence. Alternatively, a hypothesis may appear weak, but its status will be unclear until a stronger alternative is identified. For example, a study of a declining San Joaquin kit fox population began as an attempt to determine whether toxic chemicals were the cause. Its conclusion that they were not the cause became more convincing when a strongly supported alternative cause, predation, was identified (U.S. EPA, 2009a).

Because of the complexity of ecological systems and their responses to interventions, the development of conceptual models is essential to planning an ecological assessment (U.S. EPA, 1998). Conceptual models convey the processes and entities that link sources with effects on endpoints. The links can be used to guide the development of evidence for the WoE process. In risk assessments and other predictive assessments, questions can be asked such as, what sequence of events must occur following the proposed action before the endpoint effect can occur? For example, when assessing the risks from input of phosphorus to a lake, if an endpoint is fish kills due to low dissolved oxygen, the WoE should consider evidence for the occurrence of algal blooms, respiration and decomposition, and evidence against the occurrence of mixing because they all can influence dissolved oxygen (Figure 3-3). Similarly, when assessing the cause of an observed fish kill, if low dissolved oxygen is a hypothesized proximate cause, evidence of phosphorus input, algal production, decomposition, and the absence of mixing are supporting evidence that could be generated and evaluated. An approach for applying WoE to conceptual models is provided by the OECD (2013) guidance on adverse outcome pathways (Box 3-3).

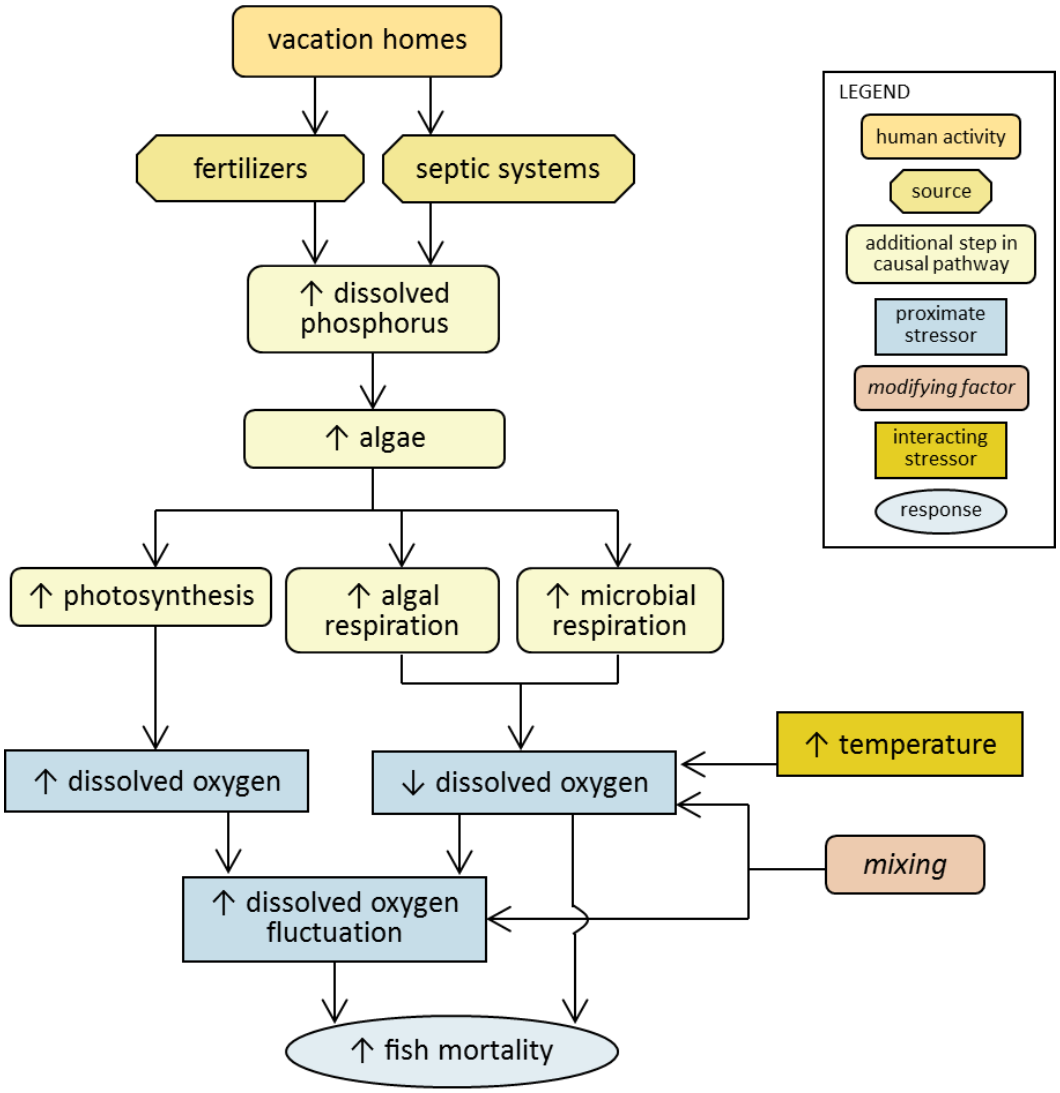


Figure 3-3. Conceptual model for a hypothetical ecological risk assessment of the relationship of phosphorous releases from a vacation home development to the risk of fish kills in a lake (graphic by Kate Schofield, using the conventions in CADDIS, at http://www3.epa.gov/caddis/).

Box 3-3. Weighing Evidence for Adverse Outcome Pathways

Adverse outcome pathways (AOPs) are representations of linkages between molecular initiating events and adverse outcomes measured at levels of biological organization considered relevant to risk assessments (Ankley et al., 2010). Although a goal of AOP development is to quantitatively implement them and predict relevant effects, currently they are primarily a form of hazard identification represented as conceptual models. For example, an AOP might link the binding of a chemical with a receptor in larval fish through a series of steps to failure of swim bladder inflation and then to reduced survival and perhaps to reduced population production (Villeneuve et al., 2014). As with other hazard identification exercises, WoE can be applied to determine the most likely hazards posed by a chemical and also to reveal gaps and weaknesses in the evidence. The OECD (2013) has recommended using Hill's considerations, tailored for AOPs, to evaluate key events, key event relationships, and overall AOPs. The tailored considerations are biological plausibility, essentiality, and empirical evidence. For each consideration, definitions of high, moderate and low confidence are provided. For example, moderate confidence in biological plausibility is defined as "the key event relationship is plausible based on analogy to accepted biological relationships but scientific understanding is not completely established." These OECD standard weights and definitions are equivalent to the standard findings and scores in Table 5-2, but types of evidence are not separately evaluated. Example cases of WoE are provided in the OECD (2013) guidance and derivative publications (Becker et al., 2015; Villeneuve et al., 2014). Subsequently, another WoE approach for AOPs was developed and demonstrated using numerical scores that were aggregated in a linear additive fashion into an overall WoE, a procedure analogous to multi-criteria decision analysis (Collier et al., 2016).

The planning of an assessment should include consideration of uncertainty. However, in the context of WoE analyses, uncertainty is one component of a set of complex issues related to confidence in the results (Section 9). As the OECD found when applying WoE to adverse outcome pathways, analysis of qualitative uncertainty is redundant with WoE (Becker et al., 2015). Conventional statistical measures such as distribution functions, standard deviations and confidence intervals are used to express the variability and uncertainty that appear as scatter in the data or scatter in modeling results. Many sources of uncertainty are unquantified or unquantifiable, however, and conventionally are simply listed. WoE can address this wider range of qualitative issues that determine confidence in the results—not just how closely the curve fits the points, but does the curve represent a causal relationship, and how relevant is the relationship to the case? A formal WoE method is a means to engage with these issues more clearly and consistently. Ideally, each quantitative result would have a conventional statistical estimate of uncertainty or variability and associated qualitative weights.

All assessments require professional judgments by assessors, but WoE makes the judgments more transparent. Clearly indicating where the assessors' judgments end and those of the decision maker begin is critical. It is clear that the decision maker, after conferring with assessors and potentially with stakeholders during the planning phase, specifies the topic and scope of the assessment and the decision to be made (U.S. EPA, 1998). It is also clear that, in the end, the decision maker determines what action will be taken. Judgments regarding the scientific evidence made between these management decisions are generally considered to be in the purview of science and are made by assessment scientists to avoid any potential political biases of decision makers (NRC, 1983). A point of potential ambiguity arrives at the conclusion of the assessment. A decision maker might request only a compilation and summary of the evidence and then decide, for example, whether the stream is impaired, which cause is the most likely or what concentration is the best benchmark value? The boundary between assessment and management, therefore, should be made clear in the planning phase because it determines the ultimate product of the assessment.

Because professional judgments play such a prominent role when weighing evidence, avoiding making the judgments in isolation is essential, particularly for less experienced assessors and for unconventional assessments. Support in professional judgment can be provided during the assessment by collaboration within the assessment team or afterwards by internal and external peer reviewers.

In the EPA, the planning of assessments includes the development and approval of a Quality Assurance Project Plan (QAPP). Although quality assurance (QA) and weighing evidence are not the same, WoE includes the screening of studies and the evaluation of evidence to weight it with respect to relevance and reliability, which encompasses evaluating the quality of input data. To avoid duplication of effort, an explicit WoE process can be used to meet the requirement for a QAPP (Box 3-4 and Section 9).

Box 3-4. Data Quality Assurance and Weight of Evidence

Quality assurance (QA) and WoE both consider the quality of scientific information and thus are related. The QA process determines whether the quality of environmental data and information supporting EPA decisions are appropriate for their intended uses (U.S. EPA, 2008). It includes documenting the process for ensuring information quality, including data generation, data analysis, and assessment methods. Thus, the development of a QAPP should encompass all aspects of the assessment planning process, including methods for WoE (U.S. EPA, 2002a). WoE contributes to QA because QA for assessments "involves a 'weight-of-evidence' approach that considers all relevant information and its quality" (U.S. EPA, 2002b). Therefore, during planning, QA should determine what WoE method is acceptable, and during the assessment, weighting determines whether the quality of information is acceptable. A formal WoE process should provide the rigor and transparency needed to meet QA requirements. Assessors should read the QA guidance and check with their organization's QA official to ensure that they properly merge QA and WoE.

3.3. Results and Transition

In this section we have dealt with determining the role of WoE in the assessment by integrating WoE considerations into the planning and problem formulation process. That is, we have identified one or more assessment problems with multiple pieces of evidence to be weighed, we have identified the WoE approach to be used, and we have determined how it relates to other aspects of the assessment. We are ready for the first step in the WoE framework, assembling the evidence.

4. ASSEMBLING EVIDENCE

4.1. The Process for Assembling Evidence

The success of WoE depends on identifying or generating useful evidence (Figure 4-1). Evidence is information that can be used to make an inference. For risk assessments and causal assessments, useful evidence principally includes information about causality (i.e., a relationship between the exposure and response and either an estimated exposure level or a response level). Exposure levels are used to solve the exposure-response relationship to estimate a future response to the potential exposure (e.g., to estimate the frequency of mortality in a conventional risk assessment) or plausible effects of the exposure (e.g., to determine whether the level of exposure was sufficient to cause the observed effect in a causal assessments; Figure 4-2). Response levels considered thresholds for unacceptability (e.g., 10-percent mortality) are used to solve the exposure-response relationship to estimate the exposure level that will protect against that unacceptable response (criteria assessments) or the exposure level that would be required to cause an observed effect (causal assessments; Figure 4-2). Exposure and response information can occur separately in a condition assessment. For example, evidence that the abundance of an endangered species is declining is sufficient for a condition assessment to prompt a causal assessment. Similarly, accumulation of a chemical in fish can be sufficient evidence of condition to prompt a risk assessment. In addition to information about exposure, response and the relationship between them, information about environmental conditions such as habitat structure or water chemistry also might be required to complete the information that constitutes a piece of evidence.

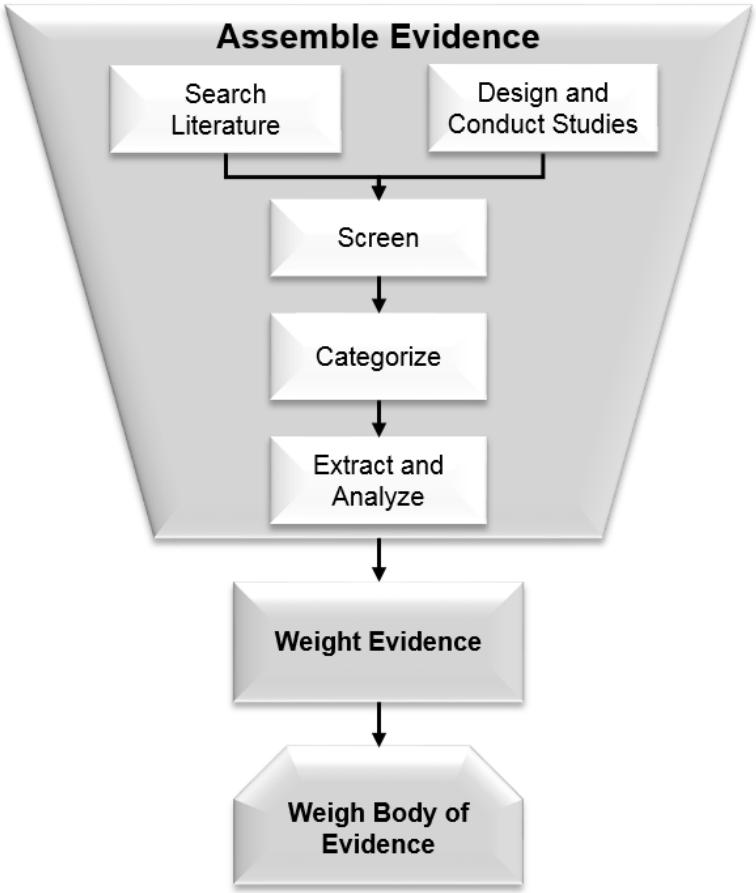


Figure 4-1. An elaboration of the process for assembling evidence, the first step in WoE.

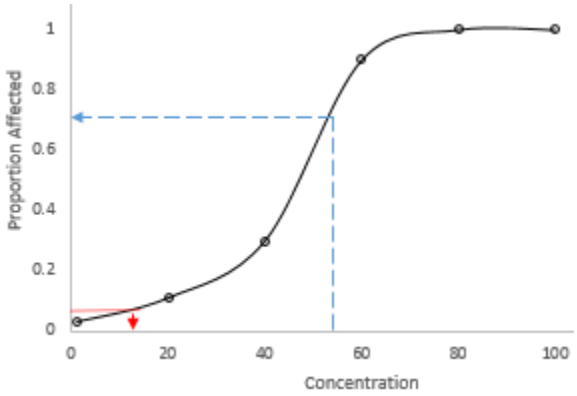


Figure 4-2. An exposure-response relationship (black curve) alone is evidence that the measured chemical can cause the effect. With a concentration, it can provide evidence of the level of effect that is expected (blue dashed arrow). With a level of effect, it can provide evidence of the concentration that would cause that level of effect or, in benchmark derivation, a concentration that would prevent greater effects (red dotted arrow).

Data alone are not evidence because evidence should indicate a hypothesized spatial, temporal, or causal relationship or the absence of a relationship. For example, contaminant concentrations are related to a particular place that has been, or potentially will be, exposed to some source and to reference concentrations from sites that are not exposed to the source. Concentrations without those relationships do not constitute evidence.

In sum, evidence usually requires more than one type of information. Recognizing this, Hope and Clarkson (2014) developed the term "evidence group" to describe the combination of an exposure-response relationship, information concerning environmental conditions that influence the relationship and either an exposure estimate or a response level. Thus, as this document discusses evidence, note that the discussion typically refers to a set of related bits of information that together constitute evidence.

4.2. Searching Literature and Assembling Evidence

A systematic literature review can provide more complete information than an informal review and can reduce the perception of bias in data selection (NRC, 2014). Recognition of the importance of performing reviews systematically has prompted development of several formal methods for systematic reviews (Box 4-1). Recent peer reviews of high-profile EPA assessments such as the stream and wetland connectivity report have called for better documented and systematic literature reviews (SAB, 2014). The method for literature review should be described in the analysis plan (U.S. EPA, 1998).

The EPA has developed the Health and Environmental Research Online (http://hero.epa.gov) database to document the literature searches and screening processes for ISAs and IRIS assessments. It provides transparency in literature searching, screening, and sorting processes and makes the results available to the public.

Box 4-1. Systematic Review

Systematic review is an approach for reviewing published scientific evidence based on a formal search protocol that provides a reasonably complete set of relevant studies that has been screened and reviewed in a consistent and replicable manner (Bilotta et al., 2014). The goal of systematic review methods is to ensure that the review is complete, unbiased, and transparent.

Systematic review was developed as a tool for integrating results of clinical trials in evidence-based medicine. The most prominent example is the Cochrane Collaboration, an organization that reviews evidence of the efficacy of medical treatments (Higgins and Green, 2011). The Campbell Collaboration (http://www.campbellcollaboration.org) extends systematic reviews to education, crime and justice, social welfare, and other issues. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) provides "an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses" (http://www.prisma-statement.org/). More recently, systematic review has been adapted to mammalian toxicology, epidemiology, and human health risk assessment. Examples include the industry-funded Evidence-Based Toxicology Collaboration (http://www.ebtox.com), the foundation-funded *Navigation Guide* (Woodruff and Sutton, 2014), and the U.S. National Toxicology Program's Office of Health Assessment and Translation method (Rooney et al., 2014). Another systematic review system has been developed for environmental management (CEE, 2013). None of the specific existing methods for systematic review are directly applicable to weighing evidence in the EPA's ecological assessments. In particular, systematic review does not integrate heterogeneous evidence, which is typically necessary in ecological assessments. However, they provide methods for the assembly of information from the literature.

4.2.1. Search the literature

The two major components of a literature search are defining the topic and designing the search. The first step identifies the topic with sufficient specificity to enable assessors or information specialists to design a search strategy. Examples include the chronic toxicity of arsenic to marine teleost fish and the biodegradation rate of benzene in fresh surface waters. If the assessments performed by an organization are sufficiently consistent, developing a standard format and content for defining the search topic might be desirable. The variants of Population, Exposure, Comparator, Outcomes, and Study Design (PECOS) systems, developed for reviews of epidemiological or clinical studies, could serve as models for defining ecological search topics (Woodruff and Sutton, 2014; CRD, 2009).

Comparisons of systematic reviews to conventional literature reviews have revealed that the latter are often incomplete or biased (Egger et al., 2003). One reason is the lack of professional assistance with searches. Some useful strategies such as the use of wild cards and truncation to obtain all forms of a search term as well as procedures for using different databases and search engines might not be familiar. Although guidance on literature searching is available (Gough et al., 2012; Higgins and Green, 2011; CRD, 2009), it is not specific to ecological assessment and could soon become outdated due to advances in information science such as text mining techniques.

Multiple search strategies and tools can be employed. Some examples include:

- Electronic databases such as ECOTOX;
- Literature search engines such as PUBMED, Web of Science, or Google Scholar;
- General search engines such as Google or Bing;

- Hand searches of key journals and books;
- Citation tracking of literature cited in key papers, books, or reports;
- Citation searching for publications citing key papers, books, or reports;
- Contacts with key authors and experts;
- Solicitation of information from stakeholders; and
- Internal databases (if they are not business confidential).

One should maintain a search log of sources searched, date, terms, and syntax (i.e., the combinations of the Boolean operators: and, or, not) used. Doing so can make the search explicit, transparent, and potentially replicable. Downloading search results to reference management software such as Endnote or Reference Manager, with a distinct name for the results of each search, also can be helpful.

The inclusion of difficult-to-locate studies (i.e., not published in a journal, not in English, or not indexed in MEDLINE) influenced the results of medical meta-analyses, but the influence varied among fields (Egger et al., 2003). Such studies tended to be less comprehensive and to have lower methodological quality. Identifying relevant studies in an unbiased fashion, however, is important. Issues of study relevance and reliability should be addressed during screening or weighting of the evidence rather than during the literature search.

4.2.2. Screen the studies

Screening the search results and studies performed for the assessment identifies irrelevant or clearly unreliable studies for elimination. Elimination criteria should be defined in advance, to the extent practicable, to avoid perceived or actual bias. Defining irrelevance is relatively straightforward. If the topic is chronic toxicity of arsenic to marine teleost fish, a freshwater invertebrate study would be eliminated. Some issues of relevance might not be apparent in advance. For example, assessors might realize during the screening process that not all arsenic species are relevant, so they should also screen out studies of irrelevant arsenic species. Screening for reliability is generally more difficult than screening for relevance and might not be performed at all. That is, assessors might choose to include all relevant studies and leave differences in reliability for the evidence-weighting phase. Eliminating studies with some obviously unacceptable attribute, such as lack of replication or lack of controls in a test, however, could be efficient. If the weighting phase is skipped (Section 7), performing a more thorough screening and considering the elimination of marginally reliable studies is important.

Some criteria commonly used to screen studies are not so clearly related to relevance or reliability. They include language (English only, unless translation services are available); peer review; and use of a standard protocol or good laboratory practices. Standard protocols ensure consistent and interpretable results, and good laboratory practices improve methodological reliability. Nonstandard studies, however, could provide relevant results and might also have good QA. Entire categories of sources could be eliminated, particularly secondary sources (if possible, primary literature should be used as sources of information), unpublished reports, or abstracts.

Some EPA programs have specific data-screening criteria. For example, the Office of Pesticide Programs screens open literature publications using 14 criteria (U.S. EPA, 2011a), and the ecological soil screening levels for plants and soil invertebrates have 11 acceptance criteria (U.S. EPA, 2005a). Other screening criteria are assessment specific, such as the assessment of connectivity of U.S. waters, which used a logic diagram for screening literature (U.S. EPA, 2015a).

When very large numbers of studies are screened, a tiered process can be used. For example, in the ISAs for criteria air pollutants, publications are screened based on their titles, then on abstracts, and finally on a full reading of the text (U.S. EPA, 2013). Duplicate screeners can minimize errors by identifying and resolving differences.

The EPA has published guidance on determining whether information quality is adequate (see Box 3-3). This procedure is conceptually equivalent to the screening process described here, but it is not followed by evidence weighting and weighing. The screening and weighting processes together should be sufficient to fulfill the EPA's mandate for information quality review.

4.2.3. Categorize the studies

Sorting studies is generally desirable so that distinct categories of evidence can be weighed before they are integrated to reach a conclusion (see Section 6.2). The categorization depends on the type of assessment, amount and diversity of evidence, circumstances of the assessment, and preferences of the assessors.

The most common approach to categorization is to assign evidence to types based on the sorts of studies from which the evidence is derived. For example, the Oak Ridge National Laboratory scheme for ecological assessment of contaminated sites divided evidence into single-chemical toxicity tests, body burdens, ambient media toxicity tests, biomarkers and pathologies, and biological surveys (Suter et al., 2000; Suter, 1996). Other types of assessments would use different types of evidence. In general, field studies are separated from laboratory studies, mechanistic studies are separated from studies of overt effects, and studies of effects at different levels of organization are separated. When only toxicity benchmark values are used as measures of effects, evidence might be categorized by mode of exposure [e.g., whole sediment, pore water, or tissue concentrations (Integral Consulting Inc., 2013)]. An assessment of a specific site might separate evidence from the site from evidence derived from other locations.

Evidence could also be categorized by the characteristic it illustrates. For example, evidence in causal assessments can be organized by characteristics of causal relationships such as interaction and co-occurrence (see Appendix E). Classification in terms of characteristics serves to explain the implication of the evidence for the inference.

4.2.4. Derive evidence from data and general knowledge

Data are raw materials for generating evidence. Data are generated by observation or experimentation and must be obtained from the investigators or extracted from the published studies that pass screening. Data can include:

- Numerical data such as the median lethal concentration (LC_{50}) values or raw exposure-response data;
- Categorical data such as whether benchmark levels are exceeded; or
- Narrative data such as the appearance of a water body, the behavior of animals, or a researcher's interpretations.

In addition to data, generally accepted knowledge such as physical laws and biological principles (e.g., receptor binding of certain chemicals or the role of species competition in structuring communities) contributes to evidence generation. Such knowledge can determine the information sought for generating evidence and the way to structure that evidence. Knowledge also might be the evidence itself. For

example, evidence that a biotic community was exposed to a release from an upstream source is provided by the general knowledge that flow would carry a persistent and soluble chemical downstream. General knowledge might appear as facts (water flows downhill) or as a mathematical model (e.g., a transport and fate model for streams).

If possible, numerical data sets should be obtained from the journal's supplementary material files or directly from the investigator. Transcribing numbers from data tables is prone to error.

If the data to be extracted are sufficiently consistent (e.g., all are from laboratory toxicity tests), development of a data form can facilitate extraction and increase completeness and consistency of the extracted data. Particularly if quantitative analysis is planned, an electronic form can efficiently combine data extraction and data entry. Any form should be piloted with a subset of the studies.

Ensuring data quality by checking the entered data for consistency with the source is desirable. One way to accomplish this is to have two individuals separately extract the data and then compare their results. Discrepancies can be due to errors or to ambiguities in the source that might be variously interpreted.

Some processing of the extracted data could be required before they can be used. In the simplest cases, a summary statistic such as an annual average is calculated. If evidence will be weighed, additional processing often is required. Evidence (as opposed to data) generally involves some sort of relationship such as the impaired stream reach was channelized, but the reference reach was not. Evidence of a condition could involve simple observation. For example, the absence of a taxon (e.g., no fish) is evidence of impairment. In addition, pieces of evidence that will be numerically combined must be converted into the same form and units (e.g., LC_{50}s converted to mg/L).

Processing data to obtain evidence also might require adjusting for the conditions to which the evidence will be applied. For example, laboratory test data might not be comparable to field data until they have been adjusted for water properties such as pH, hardness, or dissolved organic matter levels. Adjusting for the occurrence of mixtures involves applying an additivity model or other combined effects model to the individual chemical test data. Adjusting for biology might require considering seasonal occurrence of sensitive life stages.

Derivation of evidence involves the consideration of variance. Different data sets might appear to be inconsistent, until it is recognized that their distributions overlap.

4.3. Design and Conduct Studies and Assemble the Evidence

In some cases, studies can be performed to provide evidence for an assessment. In such cases, the inferential process should determine the evidence needs and drive the research planning. A simple example is provided by the sediment quality triad [(Chapman, 1996); see Appendix D, Table D-1]. That WoE method includes standard inference rules that require reliable data for the sediment's chemical composition, toxicity, and biological community composition. If any of these types of evidence are missing or unreliable, the inferential method fails. Some EPA offices and programs specify data to be generated for assessments. A prime example is the Office of Pesticide Programs' FIFRA data requirements. For other purposes, such as Superfund remedial investigations, the data needs are determined for individual cases. The Data Quality Objectives Process was developed for that purpose (Bilyard et al., 1997; U.S. EPA, 1994).

Like data from the literature, data produced for the assessment should be screened for acceptability. The Office of Solid Waste and Emergency Response provided detailed guidance to determine if contaminant data are useable (U.S. EPA, 1992a, b). If data quality objectives are developed, they should specify

standards for data acceptability. The description of data to be generated is a conventional component of the analysis plan (U.S. EPA, 1998).

As with data from literature reviews, data generated for the assessment must be input to the assessment's data management system and analyzed to derive the evidence. Unlike data from the literature, the raw data from studies performed for the assessment always should be available for analysis. Results of those studies should be categorized into the same types of evidence as the results of literature reviews.

4.4. Summary

The process of assembling evidence is more complex than the simple phrase implies. When weighing evidence, ensuring the process begins with a complete search of the literature is essential. QA procedures should be applied to the processes of obtaining or extracting data or other information from the literature. Information should be screened for relevance and reliability. Finally, information should be analyzed and organized in a way that facilitates its use as evidence.

4.5. Results and Transition

The results of assembling evidence are pieces of evidence that are at least minimally relevant and reliable, and are organized and presented in a useful form. The next step in WoE is to assign weights to the evidence that have been assembled.

5. WEIGHTING EVIDENCE

5.1. The Process of Weighting Evidence

In most assessments, some pieces of evidence will be more influential than others. The evaluation of the evidence can be implicit, as in a narrative WoE. When a formal WoE method is used, however, weighting the evidence determines the appropriate degree of influence (Figure 5-1). By not only evaluating the evidence but also expressing the weights explicitly as scores, assessors transparently determine the influence that particular pieces of evidence will have when the entire body of evidence is weighed (see Section 6). Weights are usually assigned to pieces of evidence. However, if the pieces of evidence in a category (Section 4.2.3) are similar, the category may be assigned a weight (see Section 6.2).

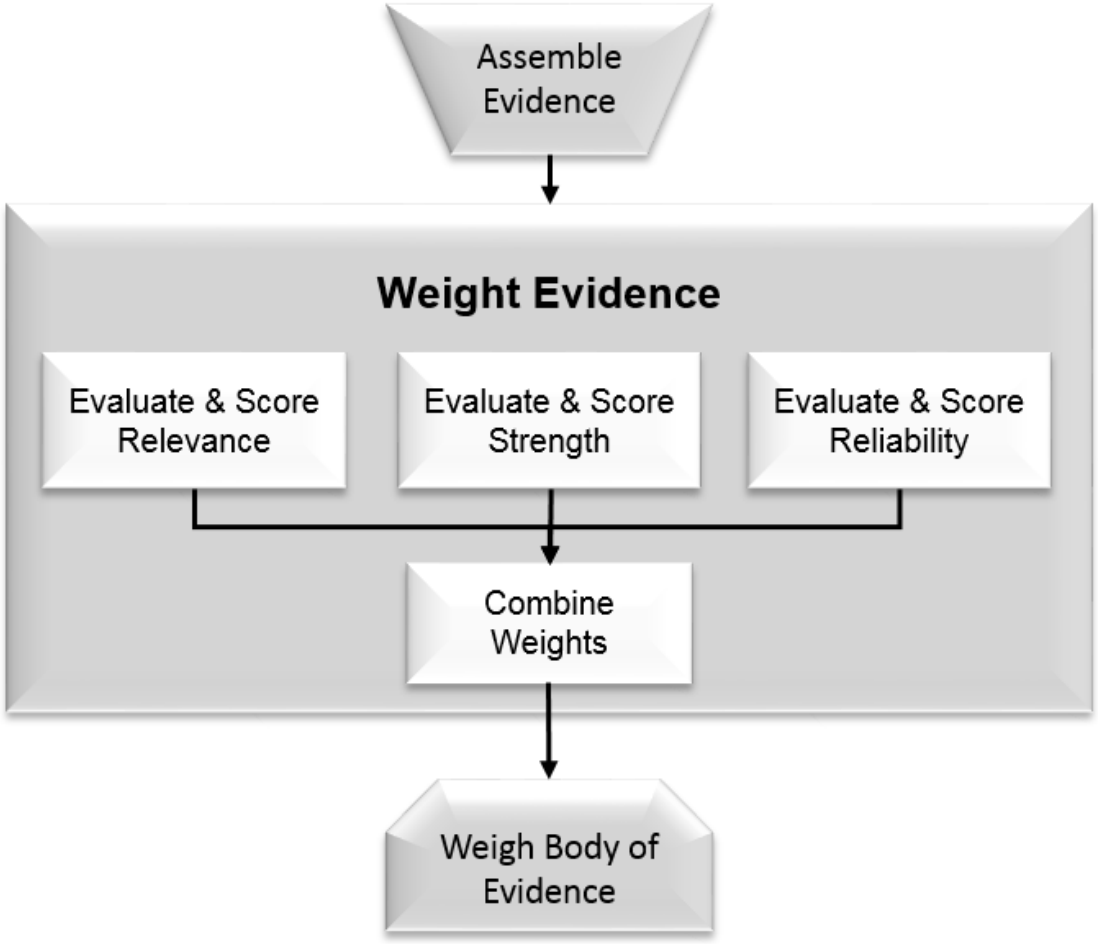


Figure 5-1. An elaboration of the process for weighting evidence, the second step in WoE.

The concept of weighting is familiar in statistics. For example, a weighted mean is the mean of study or sample values multiplied by weights, which are usually the numbers of observations. If one stream invertebrate survey of 10 sites reported a mean of 20 species and another survey of 40 sites reported a mean of 30 species, the true mean species richness across studies is found by multiplying each mean by the number of sites, adding the resulting products and dividing by 50. By weighting, the mean is shown

to be 28, not 25. Statistical weighting is an important tool of WoE for deriving quantities (see Section 8 and Appendix B) but is not considered further in this section.

The qualitative weighting of evidence has two components: evaluation and scoring. Evaluation of evidence determines the weight—how influential a piece of evidence should be, preferably based on defined properties. Scoring uses symbols to formalize the results of the evaluation. The weight is conceptual and the scores are symbolic. For example, a piece of evidence may be evaluated as moderately relevant, so the score for relevance is ++.

Many uses of WoE have not involved scoring. In these cases, either no results of evaluation are expressed or the results are expressed by descriptions. When assessors score the evidence, they are compelled to clarify the evaluation for the reader and themselves. In addition, scoring helps assessors provide clearly weighted evidence for the next step, the integration of evidence (see Section 6).

The approach presented here for qualitatively weighting evidence has been used in assessments of specific causation in the *Stressor Identification Guidance* (U.S. EPA, 2000) and in the Causal Analysis/Diagnosis Decision Information System (CADDIS, http://www3.epa.gov/caddis) and its applications. It has been used in an assessment of general causation for major ions as a cause of the impairment in stream communities (U.S. EPA, 2011b) and has been adapted to integrate evidence in a watershed risk assessment (U.S. EPA, 2014a). Below, the scoring system is presented first and then used to illustrate the evaluation of evidence.

5.2. Scoring Systems

A generally useful qualitative scoring system uses the symbols, +, − and 0, to represent evidence that, respectively, supports, weakens, or has no effect on the credibility of a hypothesis. More symbols represent greater weight. For example, the *Stressor Identification Guidance* and the CADDIS system for ecological causal assessment use the following scores.

+ + +, − − −	Convincingly supports or weakens
+ +, − −	Strongly supports or weakens
+, −	Somewhat supports or weakens
0	No effect (neutral or ambiguous)
NE	No evidence

The interpretation of plus or minus symbols is reasonably intuitive, and they have been used in other WoE systems (Rooney et al., 2014; Higgins and Green, 2011; Fox, 1991; Susser, 1986). These scoring symbols are quite general. They can be applied to a simple quantifiable property like strength of association or to a complex qualitative property like study design. Their regular use could result in clear and consist expression of WoE results. Nevertheless, other scoring systems might be more useful in particular applications.

Most other scoring systems used in environmental assessments only distinguish different degrees of support for a hypothesis. Some evidence, however, is clearly contrary to a hypothesis or has no effect. When a scoring system does not include those possibilities, a separate step must follow scoring to distinguish the logical implications of evidence [e.g., the evidence for the hypothesis is ++ for reliability and +++ for strength but it is contrary to the hypothesis (Johnston et al., 2002)]. The use of 0 and − symbols as well as + makes the logical implication of the evidence part of the weighting process rather than an afterthought.

Symbols are preferable to numerical scores because their use implies that they cannot be numerically combined. Two strongly supporting laboratory tests (++ and ++) are not equal to four somewhat supporting field tests (+, +, +, +). For a test result, a − score for study design and a + score for replication of the test do not sum to 0, because they are not commensurable. They simply signify different results for the different qualitative properties. Adding numerical scores generates a number with no units that signifies no quantity in particular.

These symbols also facilitate interpretation of bodies of evidence. When scanning a WoE table (see Section 6.2), seeing patterns in the frequencies of +, − and 0 symbols that indicate which hypotheses are supported by the weight of evidence is easier than if words or numbers are used to score evidence.

The system described above with three types of weight (positive, negative, and none) and three levels of weight (low, medium and high) has been found useful in environmental assessments, but more or less discrimination might be useful in some assessments. Binary (+/−) scores such as accept/reject and consistent/inconsistent have been recommended because more complex systems could be confusing or overwhelming (Hope and Clarkson, 2014). Alternatively, studies of survey responses have shown that people can distinguish five to seven possible response levels as in the Likert scale [e.g., very low, low, medium, high, very high (Dawes, 2008)]. Hence, the complexity of the scoring system can be adapted to the assessment and the desired degree of discrimination.

5.3. Properties to Be Weighted

Various properties of a piece or type of evidence can contribute to the degree of influence that it should exert. The specific properties fall within three general properties: relevance, strength, and reliability (defined in Box 5-1, Box 5-2, and Box 5-3).

Relevance (Box 5-1) includes biological, physical/chemical, and environmental aspects. It is at least minimally ensured when the assembled evidence is screened. Screening might be sufficient, and no further weighting of individual pieces of evidence for relevance may be necessary. For example, when assessing risk to black bears, if the only available mammalian toxicity tests are for laboratory rats and mice, the relevance of the tests cannot be distinguished because the relationships of rat and mouse sensitivities to bear sensitivity is unknown. In many cases, however, differences in study relevance are important to consider.

A strong signal is better differentiated from noise than is a weak signal, so a strong signal should be given more weight. Strong evidence shows (1) a large magnitude of difference between a treatment and control in an experiment or between exposed and reference conditions in an observational study, (2) a high degree of

Box 5-1. Relevance of a Piece or Type of Evidence

The relevance of a piece or type of evidence is the degree of correspondence between the evidence and the assessment endpoint to which it is applied.

Biological relevance—correspondence among the taxa, life stages, and processes measured or observed and the assessment endpoint (e.g., a *Daphnia* acute lethality test does not correspond well to insect life-cycle survival and fecundity, so relevance may be low).

Physical/chemical relevance—correspondence between the chemical or physical agent tested or measured at the studied site and the chemical or physical agent constituting the stressor of concern (e.g., pyrene may be used to represent a polycyclic aromatic hydrocarbon mixture, but relevance may be low).

Environmental relevance—correspondence between test conditions and conditions at the assessed site or the environmental conditions in a studied system and conditions in the region of concern (e.g., a pond mesocosm does not correspond well to a stream for many environmental parameters, so relevance may be low).

association between a putative cause and effect, or (3) a large number of elements in a set of evidence (see Box 5-2). Strength is a property of the study results, not the type of evidence or study method. The metrics for strength of evidence are familiar. Magnitude is typically represented by absolute and relative differences (e.g., body burdens were 20 times higher than at reference sites). Association is typically represented by correlation coefficients (e.g., Pearson's *r* for the correlation of emission rate and ambient concentrations was 0.8) or slopes (e.g., the regression of species richness on dissolved oxygen had a slope of 7). Number is commonly represented by the number of elements or frequency of occurrences (e.g., a fish kill has occurred at snowmelt every year for the past 6 years). Although strength occurs less frequently in WoE systems than relevance and reliability, some weight only strength (Chapman, 2007).

Box 5-2. Strength of a Piece or Type of Evidence

The strength of a piece or type of evidence is the degree of differentiation from control, reference, or randomness [modified from Norton et al. (2014)].

Magnitude—degree of difference between the amount of response at affected sites and at reference sites or in treatments and controls, between degrees of exposure or other relevant differences in the evidence, most commonly expressed as a difference between means or a ratio of means.

Association—degree to which variation in a variable representing a cause explains variation in a variable representing an effect, most commonly expressed as a correlation coefficient.

Number—the number of elements of a set of evidence (e.g., of symptoms or overt effects in a response or of steps in a causal pathway) that are reported to be observed or the number of occurrences.

Strength metrics lend themselves to standardization. For example, correlation coefficients were calculated for several associations in regional field data to determine the WoE for major ions as a cause of the extirpation of invertebrate genera and for potential confounding by other variables (U.S. EPA, 2011b). Standard scores were developed for the strength of correlations based on the authors' experience with correlations of parameters in surveys of physical, chemical, and biological properties of streams (Table 5-1). In this example, a correlation coefficient (r) between 0.25 and 0.75 is supportive but is not assigned an extra plus for strength. Because $r > 0.75$ is considered relatively strong for a correlation between a water quality measure and a biological response from a regional data set, it is given a second plus. No field correlations are believed to be convincing evidence, so none receive three + or − signs. Consistency and reasonableness of the scores are more important than the precise values chosen for the cutoffs, when they are used to compare alternative hypotheses. Documenting scoring criteria in advance also reduces the opportunities to bias the scoring relative to scoring performed without previously defined cutoff values.

Standard scores based on strength and logical implication are provided for evidence in causal assessments in the CADDIS system. The examples in Table 5-2 are based on standard alternative possible outcomes (findings) for each of 16 types of evidence and the interpretations of those outcomes in terms of the degree to which they support a potential cause. These scores show that different types of evidence have different outcomes with different implications for a hypothesis. The CADDIS table was developed for causal assessment in aquatic ecosystems. This approach to standard scoring is useful if the types of evidence and their findings are conventional. Other applications of this approach could have different types of evidence, possible outcomes, and interpretations.

Table 5-1. Weighting the strength of correlations (absolute value of *r*) and noting the logical implication—an example for evidence from stream biological surveys (Cormier and Suter, 2013; U.S. EPA, 2011b)

Assessment	Logical Implication and Strength	Score
The sign of the correlation coefficient depends on the relationship. For toxic relationships such as the correlation between conductivity and number of Ephemeroptera, the sign should be negative. Weak or positive correlations weaken the case for that candidate cause.	$\lvert r \rvert > 0.75$	+ +
	$0.75 \geq \lvert r \rvert \geq 0.25$	+
	$0.1 < \lvert r \rvert < 0.25$	0
	$\lvert r \rvert \leq 0.1$	−
	r has the wrong sign	− −

Table 5-2. Table of standard scores for 3 example types of evidence out of 15 types in CADDIS (http://www.epa.gov/caddis/si_step_scores.html). Each type of evidence is explained in its own CADDIS page.

Type of Evidence	Finding	Interpretation	Score
Spatial/temporal co-occurrence in site surveys	The effect occurs where or when the candidate cause occurs, **OR** the effect does not occur where or when the candidate cause does not occur.	This finding *somewhat supports* the case for the candidate cause, but is not strongly supportive because the association could be coincidental.	+
	Whether the candidate cause and the effect co-occur is uncertain.	This finding *neither supports nor weakens* the case for the candidate cause because the evidence is ambiguous.	0
	The effect does not occur where or when the candidate cause occurs, **OR** the effect occurs where or when the candidate cause does not occur.	This finding *convincingly weakens* the case for the candidate cause because causes must co-occur with their effects.	− − −
	The effect does not occur where and when the candidate cause occurs, **OR** the effect occurs where or when the candidate cause does not occur, and the evidence is indisputable.	This finding *refutes* the case for the candidate cause because causes must co-occur with their effects. Because the evidence is indisputable, other evidence need not be assessed.	R
Laboratory tests of site media	Laboratory tests with site media show clear biological effects that are closely related to the observed impairment.	This finding *convincingly supports* the case for the candidate cause.	+ + +
	Laboratory tests with site media show ambiguous effects, **OR** show clear effects that are not closely related to the observed impairment.	This finding *somewhat supports* the case for the candidate cause.	+
	Laboratory tests with site media show uncertain effects.	This finding *neither supports nor weakens* the case for the candidate cause.	0
	Laboratory tests with site media show no toxic effects that can be related to the observed impairment.	This finding *somewhat weakens* the case for the candidate cause but is not strongly weakening because test species, responses, or conditions might be inappropriate relative to field conditions.	−
Symptoms	Symptoms or species occurrences observed at the site are diagnostic of the candidate cause.	This finding is sufficient to *diagnose* the candidate cause as the cause of the impairment, even without the support of other types of evidence.	D
	Symptoms or species occurrences observed at the site include some but not all of a diagnostic set, **OR** symptoms or species occurrences observed at the site characterize the candidate cause and a few others.	This finding *somewhat supports* the case for the candidate cause, but is not strongly supportive because symptoms or species are indicative of multiple possible causes.	+

Type of Evidence	Finding	Interpretation	Score
	Symptoms or species occurrences observed at the site are ambiguous or occur with many causes.	This finding *neither supports nor weakens* the case for the candidate cause.	0
	Symptoms or species occurrences observed at the site are contrary to the candidate cause.	This finding *convincingly weakens* the case for the candidate cause.	−−−
	Symptoms or species occurrences observed at the site are indisputably contrary to the candidate cause.	This finding *refutes* the case for the candidate cause.	R

Properties of evidence that suggest the evidence is more reliable and should be given greater weight are listed in Box 5-3. Additional properties may be applicable in particular cases. Although scoring numerous properties for every piece of evidence could be burdensome, doing so would provide completeness and transparency for the weighting process. An alternative is to choose one or a few properties to be weighted that are judged most important or most likely to discriminate the various pieces of evidence. For example, if all evidence is consistent with prior knowledge (as is usually the case), consilience need not be scored. This approach—scoring the most important component properties of reliability—is generally useful. The least burdensome, but also least transparent, approach is to integrate the component properties of reliability implicitly for each piece of evidence and assign an overall reliability score.

Of these 11 component properties of reliability, the most attention has been devoted to study design. Developing a checklist of design features for each commonly used type of evidence is advisable. For example, Batley et al. (2002) developed a checklist of considerations for laboratory and field studies of sediment chemistry, toxicology, and community structure.

Box 5-3. Reliability of Evidence

Reliability consists of inherent properties that make evidence convincing [modified from Norton et al. (2014)].

Design and execution—evidence generated with a good study design that is well performed is more reliable.

Abundance—evidence from more numerous data is more reliable because it reflects greater replication or resolution.

Minimized confounding—evidence is more reliable when the sampling design or analysis controls extraneous correlates.

Specificity—evidence (e.g., a symptom or set of symptoms) specific to one cause or a few related causes is more reliable.

Potential for bias—evidence from a study that is not funded by an interested party, is not produced for advocacy, and is not produced by an investigator with conflicts of interest is more reliable.

Standardization—a standard method decreases the likelihood that the evidence is biased or analyses are inaccurate.

Corroboration—using models, indicators, or symptoms that have been verified by many studies and are accepted technical practice can greatly increase reliability.

Transparency—complete description of methods and inferential logic and availability of data for reanalysis provide the means to check the results and are presumed to increase reliability by reducing the likelihood of hidden faults.

Peer review—an independent peer review increases reliability of a source of information.

Consistency—the degree to which evidence does not vary in repeated instances within a study (e.g., across years, locations, sampling teams, or methods) is an indicator of reliability of a piece of evidence; when weighting types of evidence, consistency among studies of the same type is an indicator of reliability of the type.

Consilience—evidence shown to be consistent with scientific knowledge and theory, particularly with respect to underlying mechanisms, is more reliable.

Reliability is less if a study or a body of work appears biased. The potential for bias in scientific publications, although controversial, is well documented (Suter and Cormier, 2014; McGarity and Wagner, 2008). One clearly important source is publication bias, the disinclination of authors to submit and editors to accept studies with negative results (i.e., no effect is found). This bias can be detected by statistical analysis if the number of studies is sufficiently large (Ferguson and Brannick, 2012; Hunter and Schmidt, 2004). Other sources of bias can be detected only by comparing the results of studies from different sources. In particular, numerous reviews have identified different results from industry-funded studies and studies funded by neutral sources (Suter and Cormier, 2014; McGarity and Wagner, 2008). The most conspicuous example in ecological risk assessment is the difference between industry-funded and other studies and reviews concerning the teratogenicity of atrazine (Ralof, 2010). Such patterns of biases are not used to exclude data, but potential sources of bias should be identified, if possible, as they help interpret the body of evidence (NRC, 2014). For example, if results of industry-funded studies differ from government- or foundation-funded studies, they all must be carefully reviewed. If the cause of the difference cannot be determined, the reliability of the whole body of evidence may be down-weighted.

Biases may also result from personal ideologies, from a desire for publicity, or other reasons. These more personal sources are more difficult to detect but may show up in non-scientific writings.

In general, secondary sources should be screened out (Section 4.2.2), but if information from secondary sources is used, it cannot be considered highly reliable. The creation of secondary sources is an opportunity to introduce errors and biases in the extraction, analysis, and interpretation of data. Bad citing practices have been reported in the ecological literature, resulting in unsubstantiated conclusions (Sanz-Martin et al., 2016; Tood et al., 2010; Todd et al., 2007). Hence, if primary sources are unavailable, information may be taken from secondary sources but should be considered to have weak reliability. Cases in which a secondary source identifies and corrects an error in a primary source may be exceptions.

In addition to relevance, strength, and reliability, the type of evidence has sometimes been a weighting consideration. In particular, some assessors give more weight to field biological survey results than to other evidence (Chapman and Anderson, 2005). However, some field surveys have low relevance because they do not address the endpoint, do not address the sensitive taxa or other sensitive entities, or measure an irrelevant or insensitive response. In addition, field surveys might be so poorly designed or executed that the results are unreliable. For example, field data were given priority in a contaminated sediment study even though the field survey had "many limitations," including no appropriate reference sites (McPherson et al., 2008). Weighting the properties of the evidence is preferable to assuming that one type is always weightier.

Statistical analysis also has been a consideration in WoE. For example, a criterion for study evaluation in ISAs is "Are the statistical analyses appropriate, properly performed, and properly interpreted?" (U.S. EPA, 2013). However, the appropriateness of the analyses reported in a publication is immaterial if assessors perform their own analyses. Published studies should include original data or the authors should make data available for reanalysis. At least for influential evidence, assessors should consider reanalyzing the data so that results are relevant to the assessment, properly interpreted, and free of statistical errors or biases in assumptions. In such cases, evaluating the statistics the authors used is not pertinent. If original data are not available, the evidence could be given a negative score for transparency. To the extent possible, improper interpretations should be corrected (e.g., $p > 0.05$ does ***not*** mean no effect occurred), and assessors should consider down-weighting evidence that has inappropriate statistics if reanalysis is not feasible (e.g., time and resources are not available).

5.4. Tables of Weights

The primary inferential tool in weighting evidence is tabulations of pieces or categories of evidence and the weights assigned to them with respect to their properties. Because the assigned weights are expressed as scores, they are called scoring tables. Scoring tables may be aggregated to the extent that the weighting process is aggregated. Table 5-3 is a basic, generic scoring table in which both the evidence and properties are aggregated. The rows could have been individual pieces of evidence, but in this table, they are conventional types of evidence. Each of the three general properties is assigned a score based on the evaluation of the evidence each type provides. In the example, when assessing a contaminant as a potential cause of effects in the field, the results of laboratory tests of the contaminant can be positive but have low relevance to the field (+), the responses in the tests can be moderately strong (++), and the reliability of the tests can be high due to standard test designs, good laboratory practices, and regular audits (+++). In this example, the combined score (overall weight) might be + (weak supporting evidence) because the low relevance of the tests could make the strength of the results and the reliability of the method irrelevant to the inference. In other cases, a highly relevant test of the species of concern may not be sufficient to overcome very low reliability (e.g., due to absence of controls). In general, a property with very low weight will have greater influence than a moderate- or high-weight property. In

other words, a bad property of a piece of evidence tends to contaminate the whole thing. These examples illustrate why qualitative WoE is not arithmetic—you can't just add up the scores. It also illustrates why weights should be explained.

Table 5-3. Generic scoring table based on conventional types of evidence, with first line hypothetically completed. The overall weight is positive but low because the test relevance is low.

Types of Evidence	Relevance	Strength	Reliability	Overall Weight
Laboratory toxicity tests of a defined agent	+	++	+++	+
Effluent toxicity tests				
Ambient media toxicity tests				
Field biological surveys				
Field biomarkers and organ or whole-body concentrations				
Field symptoms				

Table 5-4 presents an example from a general causal assessment to determine whether dissolved major ions, measured as conductivity, cause extirpation of stream invertebrate genera. This scoring table presents three types of evidence for one characteristic of causation: the sufficiency of the observed conductivity levels to cause extirpation. Equivalent tables were presented for evidence related to each characteristic of causation. The properties of evidence were treated as binary (e.g., either the evidence is corroborated or not), so no more than one symbol was applied to a property. In this case, evidence is screened for sufficient relevance, and each relevant type of evidence is given one score designating its logical implication. Strength is scored using quantitative criteria as in Table 5-1. Corroboration is scored as the most important component of reliability in this case. Inclusion of a description of the evidence aids reader understanding.

After having assigned weights to each property of evidence (relevance, strength, and reliability), and possibly to multiple component properties of relevance, strength, or reliability, combining them into an overall weight for each piece or type of evidence might be desirable. If evidence is abundant, providing a single summary weight score can facilitate the integration step. Developing a system for combining the property weights can be relatively straightforward. The system used for the casual assessment, from which Table 5-4 was taken, is designed to provide a maximum cumulative score of three + or – units, for each of the three properties. However, as with other processes in WoE, the weights could be combined by expert judgment without defined procedures or criteria to maintain flexibility.

If evidence is scored for multiple component properties, the components should be combined before the three properties are combined. Seldom are multiple component properties of relevance or strength separately scored for a piece of evidence. Reliability has at least 11 component properties that are conceptually distinct, however, and more than one can be applied to a piece of evidence (Box 5-3). In Table 5-3, only corroboration is evaluated, but if multiple component properties of reliability were judged sufficiently important to be evaluated and scored, a separate reliability scoring table would be needed.

In any case, making the weighting convincing and transparent by explaining the reasons for the scores is desirable. This explanation could be presented in the text, but a simple statement, as in the last line of Table 5-4, might be sufficient.

Table 5-4. Example scoring table: scoring types of evidence for sufficiency (U.S. EPA, 2011b). The table is a direct copy and calls out a figure that does not appear in this document.

Type of Evidence	Description of Evidence	Log[a]	Str[b]	Cor[c]
Laboratory tests of ambient waters	A test showed acute lethality to an apparently resistant species, *Isonychia bicolor*, at conductivity levels similar to its XC_{95}.	+		
Field exposure-response relationships of biological metrics	Ephemeroptera were negatively correlated with conductivity in two data sets $r = -0.61$ and -0.72 (Figs. 2b and 4b) and $r = -0.90$ in Pond et al. (2008). This highly relevant evidence was obtained independently in two separate data sets, with moderate-to-strong correlations. Exposures were in the field with native species. Removal of sites with high levels of potential confounders had little effect on the correlation.	+		+
Field exposure-response relationships of composite indices	The field observations showed that, as conductivity increases, indices of stream condition (WVSCI and GLIMPSS) decrease [Fig. 5 and Pond et al. (2008)]. Correlations were strong [$r = -0.80$; $r = -0.90$ in Pond et al. (2008)]. Results were further corroborated by U.S. EPA (2010a). Exposures were in the field with native species.	+	+	+
Field exposure-response relationships of susceptible genera	At 500 µS/cm, the capture probabilities of more than 65% of genera began to decline. Similar results were obtained with West Virginia and Kentucky data sets.	+	+	+
Summary of sufficiency: Exposure to saline waters in Appalachia is sufficient to cause the declines of genera (+) with the salts found in the region's streams. The increases in effects of conductivity are strong even when other stressors are present (+). Different analytical approaches demonstrate that ionic strength is associated with different effect endpoints in different data sets in two states (+). The evidence is consistent. The total score is + + +.				

[a] Log = logical implication of relevant evidence, [b] Str = strength, [c] Cor = corroborated evidence.
GLIMPSS = Genus-Level Index of Most Probable Stream Status, WVSCI = West Virginia Stream Condition Index.

5.5. Not Combining Weights for Properties of Evidence

Assigning scores to each property of each piece or type of evidence might complete the weighting step. That is, separate scores for the properties, rather than a single combined weight, could be carried forward to the integration step. The advantage is that the weights assigned to each property are available to assessors while they are integrating across evidence within a hypothesis as shown in Table 6-2. In such cases, the WoE table shows which pieces of evidence are relevant, which are strong, and which are reliable. One might even score biological, physical/chemical, and environmental relevance separately if relevance is a particularly important issue in the case. The obvious disadvantage is that the integration becomes more complex because the assessor must consider weights for multiple properties while integrating multiple pieces of evidence. In addition, the relative influence of the properties might be evaluated less consistently and transparently if it is not determined in a separate step.

5.6. Summary

Once the evidence has been assembled, the WoE analysis begins in earnest with the assignment of weights to the evidence. Weights are represented by a generally useful scoring system that represents both the implication of the evidence (+, −, and 0) and the weightiness of the evidence (the number of

symbols). Three properties are scored: relevance, strength, and reliability. The evidence and associated weights are summarized in a scoring table. Weights for the three properties may be combined into an overall weight for each piece of evidence or all three may be carried forward into the next step, weighing the body of evidence.

5.7. Results and Transition

The result of the weighting step is a set of pieces or categories of evidence for each hypothesis, each of which has been assigned weights indicating how relevant, strong, and reliable it is. It is recommended that the weights be expressed as qualitative scores and organized in scoring tables. A narrative should also be provided describing the evidence and explaining the weights. This weighted set of pieces or categories of evidence are integrated and interpreted in the weighing step, which follows.

6. WEIGHING BODIES OF EVIDENCE

6.1. The Weighing Process

After weights have been assigned to pieces or categories of evidence, each resulting body of evidence is weighed to determine which, if any, hypothesis is supported. First, weighted evidence for each hypothesis is integrated to form weighed bodies of evidence for the hypotheses (Figure 6-1). Second, the bodies of evidence are interpreted to determine which hypothesis the evidence best supports. Finally, if the bodies of evidence are ambiguous or discrepant, the case should be reconsidered and additional evidence may be required.

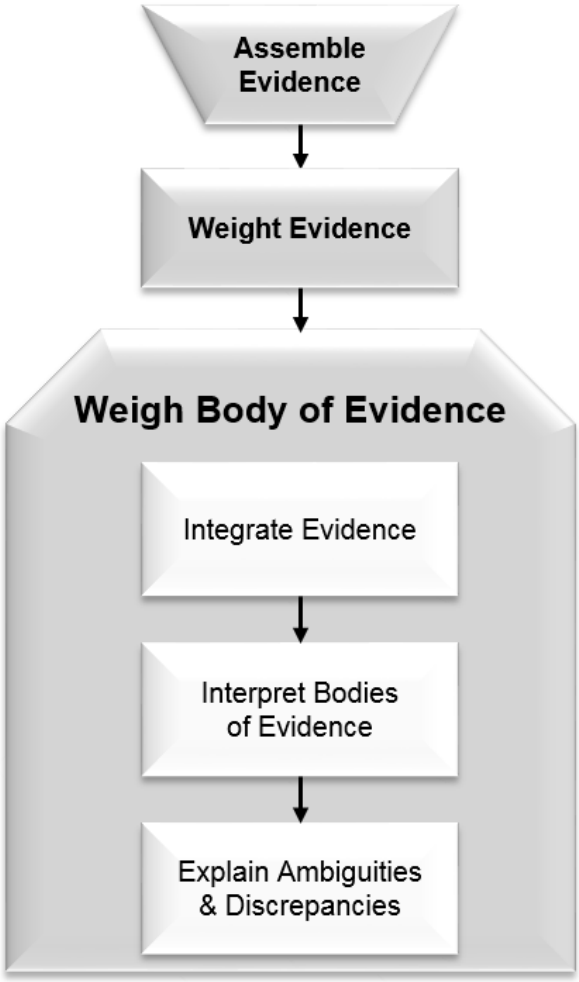


Figure 6-1. An elaboration of the process for weighing the body of evidence, the third step in weight of evidence.

6.2. Integrating Evidence

The input to the integration step is weighted pieces or categories of evidence, and the output is the weighed body of evidence for each hypothesis. The essential purpose of integration is to aggregate the evidence and associated weights into an overall weight. A secondary purpose of integration is to explain the role that the evidence serves in the inference.

1. Integration can aggregate numerous weighted pieces or categories of evidence (depending on the output of the weighting step) into a body of evidence. This purpose is typically served by deriving a combined weight for each type of evidence: evidence derived from laboratory acute tests, biomarkers, biological surveys, etc. Then, the weights are integrated across categories to weigh the body of evidence for each hypothesis.

2. Explanation can answer the question, what does the evidence do to support or weaken a hypothesis? For example, evidence from field surveys can support a hypothesized cause by showing that the cause and effect co-occur. Because WoE is often used to infer causation, the most broadly useful explanation is to associate the evidence with the characteristics of causation such as time order and interaction [Table 6-1, Table 6-4, and Table E-1 (Norton et al., 2014; Cormier et al., 2010)]. Characteristics, however, can be defined for any quality inferred by WoE such as impairment or remediation (see Appendix E). Even without defined characteristics, explanation can be performed by determining how each category of evidence relates to the hypotheses. For example, when weighing evidence for dietary bioaccumulation of a chemical in fish, concentrations in algae are evidence that the chemical occurs in the food web of the fish. If the evidence has been associated with characteristics, the sets of evidence for the characteristics can be used as categories of evidence in place of the conventional types of evidence (e.g., Table 6-4).

Evidence can be integrated in one or more steps, depending on the amount and its diversity and the circumstances of the assessment (Figure 6-2). The categories into which evidence is integrated can be types of evidence or characteristics of the quality of interest. If the pieces of evidence are few or all of one type, the body of evidence can simply be integrated after each piece of evidence is weighted (Figure 6-2a). If the differences among pieces of evidence within a category are small, weighting the pieces can be skipped and the category of evidence weighted as a whole (Figure 6-2b). When evidence is both abundant and diverse, the pieces of evidence can be weighted in each category, the categories weighed based on the weights of the pieces and the collective properties of the category (Box 6-1) and finally, the body of evidence weighed (Figure 6-2c). At the extreme, weighing a very large and diverse body of evidence can be very methodical and use multiple tiers of categories (Figure 6-2d). For example, evidence for fish acute toxicity, fish chronic toxicity, invertebrate acute toxicity, and invertebrate chronic toxicity might be separately weighted and the weights combined to weigh the evidence from all laboratory toxicity tests. Evidence from a laboratory toxicity test then could be weighed along with field tests, field surveys, and any other type of evidence to determine whether the weight of the body of evidence is sufficient to indicate that a chemical exposure is a cause of aquatic community impairment.

The degree to which the weighing of evidence is elaborated depends primarily on a tradeoff between rigor and disclosure on one hand and efficiency and simplicity on the other. The potential to confuse or overwhelm the reader with an elaborate weighing process, however, could also be a consideration. To address this problem, detailed weighting tables and text could be presented in an appendix for peer reviewers and others who need a deep understanding and a summary provided for decision makers and stakeholders.

As in the weighting step, the fundamental tool of evidence integration is a table, but this one is called a WoE table. A WoE table presents the evidence, organized in categories (types or characteristics), and the WoE scores for each evidence category and alternative hypothesis (Table 6-1). These scores come from the scoring tables created by weighting the evidence (Section 5). The same symbols are used in the weighing step as in weighting, because they express the same conceptual weights. In addition, collective properties of the body of evidence such as coherence are scored (Box 6-1). Finally, the overall weight for each hypothesis is presented. This generic table form is broadly useful, but the form of the table in any application depends on the amount of evidence, number of hypotheses, detail of the weighting, degree of

involvement of the decision maker, and other considerations. Examples of WoE tables for integrated evidence from actual assessments are presented in Table 6-2 through Table 6-4.

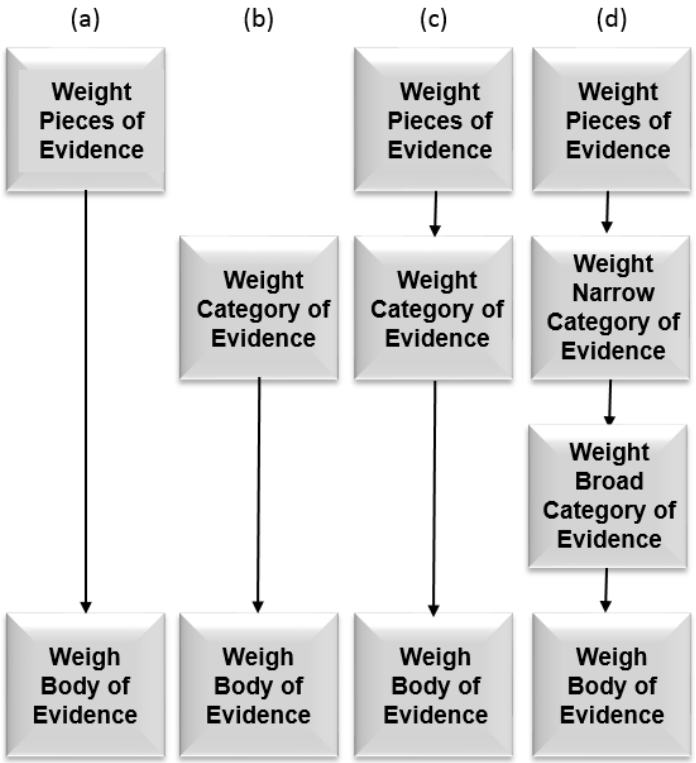


Figure 6-2. Some alternative approaches (a−d) to weighting and weighing evidence based on different approaches to aggregating evidence. Categories could be types of evidence or evidence for characteristics of the quality of interest. The choice depends on the amount and diversity of evidence, the circumstances of the assessment, and the judgments of the assessors. From Norton et al. (2014).

Box 6-1. Collective Properties of Bodies of Evidence. Modified from Norton et al. (2014).

The following collective properties of the body of evidence can be considered in addition to considering the relevance, strength, and reliability of the pieces or types of evidence. As with the properties of pieces or types of evidence, any of these could be used, depending on their relevance and utility for distinguishing among the hypotheses.

Number—more pieces or categories of evidence within a body of evidence increase the reliability, if they are consistent and are generated independently.

Coherence—logical consistency among types of evidence increases the reliability of a body of evidence, particularly when empirical and mechanistic evidence and evidence from a case and from elsewhere are coherent.

Absence of bias—consistent results from different funders and types of investigators (e.g., academic, industry, nongovernmental organization, government) are more reliable.

Diversity—evidence from more responses, taxa, life stages, or communities under more conditions is more likely to detect important exposures, responses, and relationships reliably.

Table 6-1. A generic weight-of-evidence table for *n* alternative causal hypotheses (H_1, H_2, … H_n), based on causal characteristics and collective properties of the bodies of evidence

	Combined Weight		
	H_1	H_2	H_n
Characteristics of Causation			
Co-occurrence			
Sufficiency			
Time order			
Interaction			
Specific alteration			
Antecedents			
Body of Evidence, Collective Properties			
Number			
Coherence			
Absence of bias			
Diversity			
Integrated WoE			
WoE for the hypothesis			

Table 6-2 presents, for illustrative purposes, a subset of the types of evidence and causal hypotheses from a WoE table for determining the cause of a precipitous decline in the abundance of San Joaquin kit foxes on the Elk Hills Naval Petroleum Reserve, California (U.S. EPA, 2009a). The full table occupies multiple pages. The evidence was scored using the types and standard scores in CADDIS, which eliminates the need to weight the evidence for relevance, strength, and reliability. However, it requires accepting the relevance of the standard weights and scores. One collective property of the body of evidence (coherence) is the primary basis for weighing the body of evidence. The conclusion was that predation by coyotes was the dominant cause of the decline and the evidence was convincing. The

coyotes appear to have increased on the site (co-occurrence), the cause of death of radio-collared foxes was known (evidence of exposure or mechanism), population modeling showed that predation was sufficient alone to account for the decline (stressor-response relationship), and coyote control coincided with an end to the decline (manipulation of exposure). Toxic chemicals were present and exposure occurred and was documented by analysis, but far too few foxes were significantly exposed to account for the decline. Accidents contributed but the contribution was relatively small. Note that the + for coherence of disease indicates that the evidence consistently fails to support that hypothesis, so that cause can be eliminated.

Table 6-2. Partial WoE table for alternative possible causes of the decline of San Joaquin kit foxes. Only 4 of 6 potential causes and 7 of 16 types of evidence are included. Adapted from **U.S. EPA (2009a)**.

Types of Evidence	Predation	Toxics	Accidents	Disease
Spatial/temporal co-occurrence	+	+	+	−
Temporal sequence	0	0	NE	NE
Evidence of exposure or biological mechanism	++	++	++	−−
Causal pathway	+	+	+	0
Manipulation of exposure	+	NE	NE	NE
Stressor-response relationships from simulation models	+++	−	+	−−
Coherence	+++	−	+	+

Table 6-3 provides an example of a WoE table from a risk assessment. One hazard in the Bristol Bay watershed assessment was the potential failure of a diesel fuel pipeline at a stream or river crossing, so evidence was weighed for the hypothesis that a spill at a crossing would reduce salmon production (U.S. EPA, 2014a). Unlike the generic types of evidence, Table 6-2 shows types of evidence that are specific to the hypothesis. Evidence concerning the exposure-response relationship was available from laboratory tests of dissolved or dispersed diesel fuel and from eight studies of diesel spills into streams. Exposure was estimated by modeling hypothetical spills to estimate dissolved concentrations, diesel/water ratios, and total volume spilled. The results of combining the weights across properties for each type are presented as brief narratives rather than as combined scores, but the scores for each property are combined across types of evidence. This WoE table shows the consistency of results (all are supportive). It also illustrates where the weakness in the evidence lies—studies of oil spills in streams have not characterized exposure well, as indicated by ambiguous reliability.

Table 6-4 presents a table with combined scores for assessing whether major ions in streams cause extirpation of benthic invertebrates (U.S. EPA, 2011b). The evidence is categorized in terms of characteristics of causation and is derived from the results of scoring tables for each characteristic like Table 5-4. That is, the summary score from weighting evidence for sufficiency in Table 5-4 (+++) is the sufficiency score in Table 6-4 (also +++).

Table 6-4 illustrates how explanation of the pieces or types of evidence in terms of characteristics of causation elucidates the meaning of the evidence for a hypothesis. For example, elevated concentrations of a chemical in numerous streams with impaired communities mean that the chemical and community *co-occur*, and therefore, exposure is likely. If the concentration is greater than concentrations that cause relevant toxic effects, that means that the exposure believed to be *sufficient* to cause the effect. Although such explanations of the evidence might appear self-evident to assessors, it can be an important part of the inference. Although organizing evidence in terms of characteristics serves primarily to help explain the

implications of the evidence, it can also show that a study can provide evidence of multiple characteristics. For example, a laboratory toxicity test can provide evidence of sufficiency (the level of exposure required to cause effects) and of alteration (symptomatic effects of the exposure that may be observed in the field).

Table 6-3. Summary of evidence concerning risks to fish from a diesel spill (U.S. EPA, 2014a). The risk characterization is based on weighing four pieces of evidence for different routes of exposure. All evidence is qualitatively weighted on three properties: logical implication, strength, and reliability of methods. Here, all pieces of evidence have the same logical implication: all suggest a diesel spill would have adverse effects.

Route of Exposure Source of Evidence (Exposure/E-R)	Logical Implication[a]	Strength	Reliability		Result
			Exposure	E-R[b]	
Dissolved hydrocarbons: Model/laboratory acute tests	+	+	0	+	Modeled dissolved diesel concentrations are clearly lethal to invertebrates and approximately lethal to trout.
Dissolved hydrocarbons: Model/laboratory-based standard	+	++	0	++	Modeled dissolved diesel concentrations greatly exceed the state standard.
Dispersed hydrocarbons: Diesel oil:water ratio/laboratory acute tests	+	++	0	+	Diesel oil:water ratios in the spills and in tests suggest lethality to invertebrates and trout.
All routes in actual spills: Amount spilled/observed effects	+	++	+	+	Diesel spills in other streams cause acute biological effects.
Integrated weight of evidence	+	++	0	+	The effects by four types of evidence are consistent, and the observed effects are strong. The greatest uncertainty is the relation of laboratory to field exposures.

Notes:
[a] Logical implication indicates relevance and a particular direction of the evidence (supports or weakens)
[b] E-R = exposure-response relationship

Table 6-4. Example of weighing evidence for a potential cause, major ions measured as conductivity, of the loss of macroinvertebrate genera (U.S. EPA, 2011b). The evidence is organized in terms of characteristics of causation.

Characteristic	Body of Evidence	Scores
Co–occurrence	Loss of genera occurs when conductivity is high but is rare when conductivity is low.	+++
Antecedence	Sources of the ionic mixture are present and are shown to increase stream conductivity in the region.	+++
Interaction	Aquatic organisms are directly exposed to dissolved ions. Based on first principles of physics, ionic gradients in high-conductivity streams would not favor the exchange of ions across gill epithelia. Physiological studies over the past 100 years have documented the many ways that physiological functions of organisms are affected by the relative amounts and concentrations of ions (i.e., combinations of ions that some genera do not have mechanisms or the capacity to regulate).	++
Alteration	Some genera and other response metrics and assemblages are affected at sites with higher conductivity, whereas others are not. These differences are characteristic of relative sensitivity to high conductivity.	+++
Sufficiency	Laboratory analyses report results of effects for a tolerant species, but test durations and most ionic compositions are not representative of exposure in streams. Based on field observations, however, regular increases in effects on invertebrates with increased ionic exposure indicate that exposures are sufficient.	+++
Time order	Conductivity is high and extirpation has occurred after mining permits are issued, but conductivity and biological data before and after mining began are not available.	NE
Summary of body of evidence	Five characteristics are supported and none weaken the body of evidence that increases in conductivity causes extirpation of freshwater benthic invertebrates.	Very likely

6.3. Interpreting Bodies of Evidence

Interpretation is the step in which the bodies of evidence for each hypothesis are used to determine which, if any, hypothesis is supported, and therefore, is likely to be true. Depending on the case, the interpretation of evidence could be performed by comparing alternative hypotheses or by judging whether the evidence sufficiently supports a particular hypothesis. Comparison of alternatives typically occurs in specific causal assessments (e.g., What is causing the low species richness in Jones Creek?). The hypothesis with the greatest weight of evidence is the most likely of the assessed possible causes. This is not just a matter of counting the scores. In particular, the coherence of each body of evidence with respect to the associated hypothesis is considered.

Multiple hypotheses might be likely or at least well supported. For example, both a water treatment plant effluent and stream channelization might be sufficient causes of low species richness in a stream, so both hypotheses could be accepted.

Judging a single hypothesis can be conceptualized as comparing the hypothesis against its negation (e.g., WoE for teratogen versus WoE for not a teratogen) or as a comparison of the weight of evidence for the hypothesis with a standard of sufficient evidence (e.g., a sufficient WoE for teratogenicity versus insufficient evidence). One particular hypothesis can be judged in some particular types of assessments:

(1) a general causal assessment (e.g., Does selenium cause terata in frogs?), (2) specific causal assessments for which only one potential cause is considered (e.g., Does the effluent cause the observed biological impairment?), (3) a condition assessment (e.g., Is the stream impaired?), or (4) an outcome assessment (e.g., Has the wetland recovered?).

These approaches to weighing evidence are consistent with that for the conventional standard of proof in civil legal proceedings and public policy, the "preponderance of the evidence." To meet a preponderance-of-evidence standard, a hypothesis must be shown to be more likely than its negation or than the alternative hypotheses. This standard is met by comparing the weights of the bodies of evidence for a hypothesis and its alternatives.

Hume (1748), Laplace (1812), and Sagan (1980) all made the point that extraordinary claims require extraordinary evidence. This inferential rule should be kept in mind when interpreting bodies of evidence. A hypothesis is extraordinary if there is no prior evidence for it (e.g., the agent has not previously been shown to cause the effect or has been shown to be effective only at much higher levels), or it is simply not plausible. Extraordinary evidence is very weighty evidence: highly relevant, strong, and reliable.

Interpreting the evidence for hypotheses usually involves applying a three-value logic (e.g., yes, no, maybe; true, false, uncertain; or high, low, intermediate). If the interpretation is comparative, as in determining which alternative cause is best supported by the evidence, examining a summary table of the scored evidence and identifying the likely causal hypothesis are possible. Even if the best hypothesis is not clear, proposed hypotheses that clearly are not supported and can be eliminated can nearly always be identified. For example, an assessment of the cause of a precipitous decline in smallmouth bass in the Susquehanna and Juniata Rivers, Pennsylvania, using the CADDIS WoE method and existing information was inconclusive (Shull and Pulket, 2015). It did, however, classify the 18 candidate causes into three bins: 2 likely, 8 unlikely, and 8 uncertain. These results are being used to prioritize and design subsequent research and monitoring. Similarly, when judging a single hypothesis (e.g., spilled tailings would pose a substantial risk to salmon production), the hope is for a body of evidence that clearly either supports or counters the hypothesis, but the evidence might be ambiguous (U.S. EPA, 2014a).

Some WoE systems have standard categorical outcomes defined in terms of properties of the body of evidence (U.S. EPA, 2013, 2005b). A description of the potential weights of evidence used for causal determinations (i.e., categorical outcomes for general causal assessments) of criteria air pollutants is shown in Table 6-5. In such cases, interpreting the evidence involves matching the properties of the body of evidence to one of the categories. The definitions of categories, however, could be based on the scored properties rather than narratives as in this example.

Ultimately, the interpretation of evidence is a matter of logic applied to evidence and background knowledge, preferably by multiple individuals with a range of expertise. It is not simply a matter of adding or counting scores. In particular, some pieces of evidence are conclusive alone. For example, if the effect precedes a potential cause or if an aqueous effect occurs upstream of the putative source, the cause can be eliminated even if other evidence is positive. More commonly, one piece or type of evidence will influence the interpretation of other evidence. Even WoE systems that use standard scoring criteria and arithmetic integration of numerical scores rely on logic to identify and correct results that are contrary to knowledge of the system (COA Sediment Task Group, 2008; Johnston et al., 2002). As a result, applying expert judgment in any WoE is necessary to achieve an adequate, explanatory account.

At this stage, it is desirable to limit the interpretation to the hypotheses and evidence that were defined in advance. Otherwise, "ad hoc machinations" to achieve an acceptable answer can occur (Douglas, 2012). If interpreting the evidence does not identify a clear preponderance of evidence for a hypothesis, techniques that go beyond the simple weighing of evidence can be applied, as explained in the following

section. However, it should be remembered that they violate the admonition against improvised reasoning so they are less convincing than results of weighing bodies of evidence and hypotheses defined in advance.

Table 6-5. Weight of evidence for causal determinations in the 2013 lead ISA (U.S. EPA, 2013).

Causal Determination	Evidence for Ecological and Welfare Effects
Causal relationship	Evidence is sufficient to conclude a causal relationship with relevant pollutant exposures (i.e., doses or exposures generally within one to two orders of magnitude of current levels). That is, the pollutant has been shown to result in effects in studies in which chance, bias, and confounding could be ruled out with reasonable confidence. Controlled exposure studies (laboratory or small- to medium-scale field studies) provide the strongest evidence for causality, but the scope of inference could be limited. Generally, determination is based on multiple studies conducted by multiple research groups, and evidence that is considered sufficient to infer a causal relationship is usually obtained from the joint consideration of many lines of evidence that reinforce each other.
Likely to be a causal relationship	Evidence is sufficient to conclude a likely causal association with relevant pollutant exposures. That is, an association has been observed between the pollutant and the outcome in studies in which chance, bias, and confounding are minimized, but uncertainties remain. For example, field studies show a relationship, but suspected interacting factors cannot be controlled, and other lines of evidence are limited or inconsistent. Generally, determination is based on multiple studies in multiple research groups.
Suggestive of a causal relationship	Evidence is suggestive of a causal relationship with relevant pollutant exposures, but chance, bias, and confounding cannot be ruled out. For example, at least one high-quality study shows an effect, but the results of other studies are inconsistent.
Inadequate to infer a causal relationship	The available studies are of insufficient quality, consistency, or statistical power to permit a conclusion regarding the presence or absence of an effect.
Not likely to be a causal relationship	Several adequate studies, examining relationships with relevant exposures, are consistent in failing to show an effect at any level of exposure.

6.4. Explaining Ambiguities and Discrepancies

The results of weighing bodies of evidence can be ambiguous and discrepancies among pieces or types of evidence can occur. Examples include WoE results for which:

1. The bodies of evidence for all assessed hypotheses are incoherent.

2. One hypothesis has sufficient, consistently positive evidence to support acceptance, but some other hypotheses also have some relevant and reliable positive evidence.

3. No hypothesis has predominantly positive evidence.

In ambiguous, discrepant, or nonsensical cases, one can stop the assessment process and call for more data or proceed to perform follow-on analyses to resolve problems with the existing evidence. The latter course could allow successful completion of the assessment by reconsidering the evidence or by reformulating the hypotheses. Having completed the WoE analysis, some hypotheses could have been

eliminated, but the grounds for elimination might need re-examination. Also, assessors should have become deeply familiar with the evidence. That familiarity and the assessors' background knowledge could make it possible to reinterpret the data or revise the hypotheses to explain and resolve the ambiguities. Reinterpretations, however, provide opportunities to fabricate false explanations of apparent patterns in the evidence. Formulating hypotheses to accommodate the evidence is discouraged in scientific inference because it can lead to bias or self-deception. It has come to be known as HARKing (hypothesizing after results are known). As in other judgments during the assessment process, clearly articulating arguments, avoiding over interpretation, involving multiple assessors with different perspectives, and using a clear and consistent method to build the case are essential. A generally useful approach is to (1) list the discrepancies, (2) ask whether each discrepancy could be explained by a misinterpretation of the evidence or by a misspecification of the hypothesis, (3) determine whether all discrepancies can be resolved by some combination of reinterpreting evidence and respecifying hypotheses, and (4) evaluate additional evidence relevant to the explanations.

Reinterpreting the evidence is facilitated by considering ambiguities in the data and the methods used to generate them. As presented in Appendix E, Table E-3, toxicity tests might not include sensitive species, life stages, or responses; the bioavailability or form of a toxicant might be inappropriate; the exposure durations might be too long or too short, etc. Biological surveys could be conducted in the wrong season, compared to an inappropriate reference, to measured responses or taxa that are insensitive to the causal agent, etc. Measures of exposure might miss episodic events, be taken under atypical conditions, not include the causal agents, be disjunct from the biological samples, etc. Any of the evidence could be derived from studies that are biased, poorly designed, or poorly conducted in ways that were not reported or not recognized in the weighting. Inventing explanations is not difficult—the difficulty is determining which are likely to be true.

In addition to reinterpreting the evidence, assessors might explain discrepancies by changing the hypotheses or adding hypotheses in light of the evidence. At least four strategies can be applied.

Redefine the endpoint effect. Causal relationships are often unclear because the effect is unclear. For example, determining the cause of decline in peregrine falcons depended on redefining the effect as reproductive failure associated with thin eggshells. That more specific definition steered assessors away from potential causes such as habitat loss, shooting, and egg collecting and toward toxic effects. In some cases, discrepancies might be explained by defining the spatial or temporal scope of the effect more specifically. For example, defining the effect as the occurrence of fish kills in a stream as a whole will lead to discrepancies in the evidence if kills occur only in a reach below an outfall and the dead fish drift downstream.

Reconsider sources and agents. Discrepancies in evidence could be due to sources, components of emissions, or actions that were not considered, because they were overlooked or of no concern to the decision maker or stakeholders.

Integrate causal agents or networks. Rather than being alternatives, the proposed causes of an effect could be acting jointly. Also, a proposed proximate cause might actually be an indirect cause that contributes to the true proximate cause. For example, habitat disturbance did not cause the decline in San Joaquin kit foxes on the Elk Hills Petroleum Reserve, but it apparently made the foxes more susceptible to the major cause, predation (U.S. EPA, 2009a). Such possibilities can be represented in practice by rearranging the components of a conceptual model or by combining alternative conceptual models. The conceptual model can be revised by removing the box-and-arrow combinations that are not supported by evidence and examining how the remaining boxes and arrows might link within the model or among models for different causes.

Look for patterns in the evidence. Looking for properties that the pieces or categories of evidence supporting a hypothesis share, which the contrary evidence does not share, might be helpful. This examination could be facilitated by creating a matrix of evidence versus relevant attributes of the evidence. Different results might be observed, for example, in field versus laboratory studies, lotic versus lentic studies, industry-funded versus foundation-funded studies, insect versus crustacean studies, etc.

Reinterpreted evidence or revised hypotheses could explain discrepancies in a manner that seems convincing to the assessors who developed them, but explanations improvised after the analysis to account for discrepancies might not convince others without independent evidence. Independent evidence might have been available but unused in prior weighing of evidence, because it had not been relevant. For example, if the proposed resolution of a discrepancy was low bioavailability of metals, measurements of dissolved organic matter might be used to confirm that explanation. Similarly, if low dissolved oxygen is a proposed cause of a fish kill, the survival of air-breathing organisms such as frogs and turtles is supportive (Cormier et al., 2010). Ideally, confirmatory evidence would be generated by identifying a previously unmeasured or unobserved property of the system that should occur under the revised hypothesis and then taking the measurements or observations that would establish its occurrence. If the amount of evidence and the number of possible explanations of the evidence are substantial, a formal reweighing of the evidence to address the revised hypotheses might be appropriate.

6.5. Presenting Results

The results of a qualitative WoE have two parts: the conclusion and the justification. The conclusion, in the best case, is a statement of the hypothesis that is clearly supported by the WoE (e.g., the cause is the thermal effluent, the stream is biologically impaired, or the population has recovered). Ambiguous results require more explanation (e.g., hypothetical causes 1, 3, and 5 can be eliminated, but 2 and 4 are somewhat supported). Graphical presentations can be useful, such as a revised conceptual model of the identified causal relationship or a map of the areas biologically impaired by the waste. When the assessments are sufficiently standardized, a standard reporting format could be implemented as in the International Uniform Chemical Information Database (IUCLID) dossier for assessing the toxic properties of chemicals in the European Union (ECHA, 2010).

The justification of a conclusion is a summary of the evidence and logic that support the conclusion. The justification, which will vary among assessments, typically includes:

1. Statements of overall weight of evidence for the conclusion (e.g., the evidence convincingly supports, strongly supports, or somewhat supports the cause) relative to the weights for alternatives or relative to a standard of evidence.

2. Statements of consistency or coherence (e.g., all the evidence was either convincingly or strongly supportive of the cause, except for laboratory acute tests, which had low relevance).

3. Statements of the extent to which characteristics of causation or other relevant characteristics are satisfied by the evidence (e.g., evidence was available to support all characteristics of causation except time order). Alternatively, state the extent to which standard types of evidence were available (e.g., all three components of the sediment quality triad were available). In any case, state how the body of evidence provides an explanation of the credibility of the accepted hypothesis.

4. Statements explaining why the conclusion is justified despite a low weight of the body of evidence, if necessary (e.g., even if other causes could be substantially contributing, collection and treatment of the effluent is likely to improve conditions and is mandated by regulations).

6.6. Summary

Having assigned weights to the pieces of evidence, the process of weighing evidence integrates and interprets the body of evidence to determine which, if any hypothesis has sufficient weight to be accepted. Depending on the amount and types of evidence, this may include a process of aggregating the evidence into categories and integrating the weights within categories. Once the weights have been integrated, the bodies of evidence for all hypotheses should be compared and the results interpreted. Interpretation may be obvious (e.g., only one hypothesis has positive evidence) but often it requires careful logic and judgment informed by the knowledge and experience of multiple assessors. If interpretation does not provide a conclusion, an analysis of the ambiguities and discrepancies that confounded the interpretation can suggest what additional evidence or alternative hypotheses should be considered. Finally, the results should be presented in a way that make clear the results and the process that generated them.

7. SPECIAL CASES AND ABBREVIATED PROCESSES

In some circumstances, the full three-step WoE process is not necessary or practical, but a two-step process is useful. The systematic assembly of evidence is always appropriate, but weighting might be unnecessary if all evidence is equivalent (Figure 7-1a), and weighing a body of evidence is unnecessary if only one piece of evidence is available (Figure 7-1b).

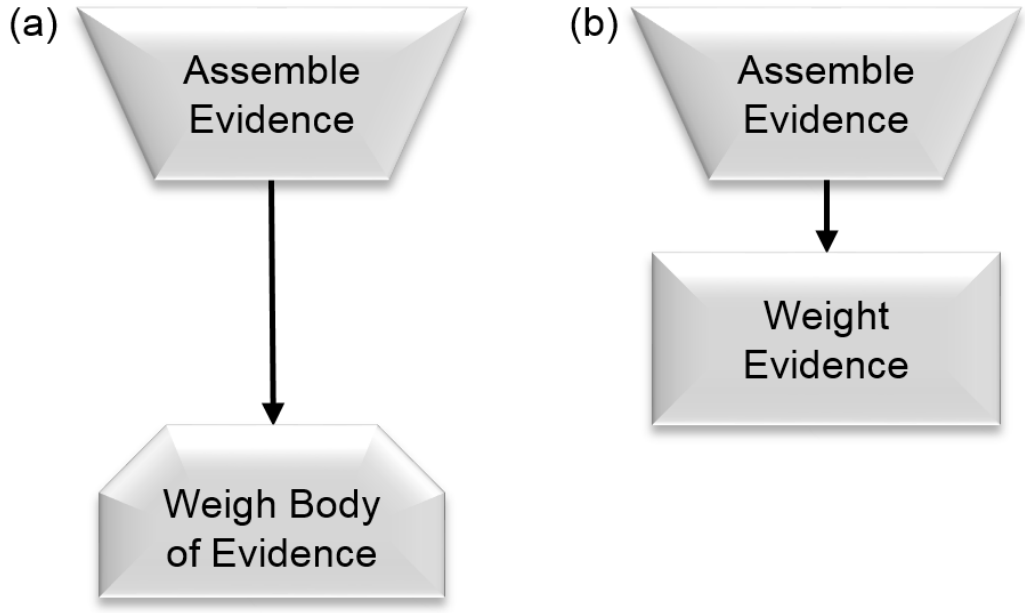


Figure 7-1. Steps in abbreviated weight-of-evidence processes: (a) skipping the weighting step when all evidence is equivalent or (b) weighting a single piece of evidence when multiple pieces are not available.

7.1. Weighing Without Weighting

Some WoE methods presume that all evidence is equal or at least that the differences are uninformative, so the weighting step (Section 5) can be skipped. For example, if all evidence is reliable and of the same type (e.g., multiple acute lethality tests performed by standard protocols), each piece of evidence will have equal influence in most cases. In practice, weighting is often skipped without determining that the relevance, strength, and reliability of the evidence are similar across pieces of evidence. Instead, if the pieces of evidence pass the screening step (see Section 3), the differences are assumed unimportant. An example of weighing evidence without weighting is the causal determination in an ISA regarding exposure to a specific air pollutant and specific effects (U.S. EPA, 2013).

A case for which distinguishing weights was unnecessary is provided by the studies used to determine that ingesting lead from lead mining and smelting caused the tundra swan kills in the Coeur d'Alene basin, Idaho (Table 7-1). The data used in the evidence were generated for the case, so they were all highly relevant; the studies were conducted by competent and reputable investigators following prescribed quality standards, so they were highly reliable; and the relationships were all strong. Because of that ideal situation, the presentation of evidence in WoE Table 7-1 is sufficient. Table 7-1 organizes the evidence in terms of characteristics of causation, rather than listing pieces or types, and briefly summarizes the evidence for each.

Table 7-1. Summary of evidence for lead as a cause of mass mortality of tundra swans in the Coeur d'Alene River Watershed (Norton et al., 2014). Based on evidence from URS Greiner Inc. and CH2M Hill (2011).

Causal Characteristic	Evidence
Co-occurrence	Swan kills occurred in Pb-contaminated lakes and wetlands and not elsewhere in the region.
Sufficiency	Mortality occurred in laboratory tests at Pb doses and body burdens observed in dead or moribund swans in the field. Consistent mortality in the field at blood Pb levels >0.5 µg/g.
Time order	No evidence—no pre-mining information on swan mortality.
Interaction	Dead and moribund swans had high blood and liver Pb levels. Pb-contaminated sediments were found in swan guts and excreta.
Specific alteration	Swans had pathologies characteristic of Pb toxicity, particularly, enlarged gall bladders containing viscous dark green bile.
Antecedents	Spills of Pb mine tailings and atmospheric deposition from smelters account for the high sediment Pb levels.

7.2. Weighting a Single Piece of Evidence

The concept of WoE was developed to combine multiple pieces of evidence in a way that gives each piece proper influence on the conclusion. When only one piece of evidence is available for an inference, the integration function of WoE does not apply. WoE, however, also leads assessors to consider the importance that should be assigned to a body of evidence and reveals the assessors' judgments to decision makers and stakeholders. This purpose of evaluating and interpreting a body of evidence can apply to a single piece. That is, explicitly evaluating the relevance, strength and reliability of a single piece of evidence can be informative.

Weighting a single piece of evidence also might provide consistent communication of the assessment process. That is, if many pieces of evidence are evaluated and scored for some hypotheses, the reader will expect to see scores for all hypotheses. If a hypothesis has only a single, stand-alone piece of evidence, scoring that one piece provides the expected consistency in weighing the evidence for all hypotheses.

In addition, weighting a single piece of evidence might reveal its inadequacy and lead to obtaining and weighing more evidence. For example, for the Patrick Bayou Superfund site, assessors intended to use modeling of toxicity to an amphipod as their primary approach to assessing risks to benthic invertebrates (Anchor, 2013). However, they found that the data were unreliable due to lack of reference data, confounding, poor correspondence to tests of other species, and unreliable polychlorinated biphenyl analyses. As a result, they adopted a WoE approach using the conventional triad of sediment quality evidence.

8. WEIGHT OF EVIDENCE FOR QUANTITATIVE RESULTS

Quantitative results of ecological assessments, such as benchmark concentrations, areas of habitat lost or rate constants, might be derived from multiple pieces of quantitative evidence. Commonly, these quantitative analyses follow the qualitative analyses that establish the hazard or other qualities to be quantified (Figure 8-1). Methods for deriving quantities by weighing evidence are discussed in Appendix B. They fall into two basic approaches: combine the quantitative evidence or choose the best quantitative evidence. Finally, weights can be assigned to the quantitative results to indicate how influential they should be.

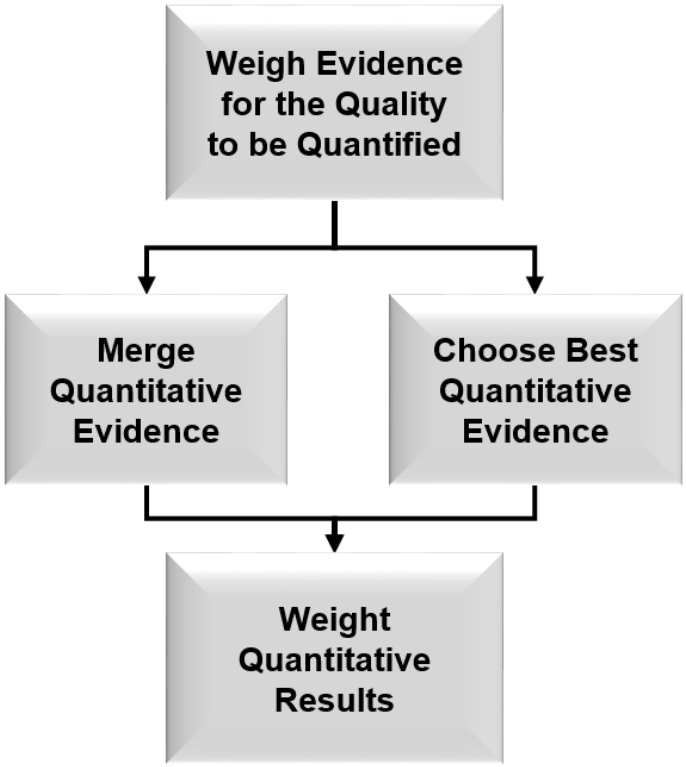


Figure 8-1. Potential steps in a process for using WoE to derive a quantitative result. Note that the top box of this process diagram encompasses the qualitative WoE process (Figure 3-2).

8.1. Weight of Evidence for the Quality to Be Quantified

Before deriving a quantity, determining what quality is to be quantified (Table 8-1) is necessary. This determination could require weighing evidence. For example, conventional risk assessments quantify, if possible, the magnitude and likelihood of a defined hazard. Environmental hazards are potential effects on an attribute of a biological entity (the assessment endpoint) resulting from exposure to an agent in particular conditions (described by the conceptual model). Similarly, criterion assessments quantify a threshold level of exposure corresponding to an unacceptable level of effect, and the effect itself is a quality that might be identified by WoE. Qualitative assessment could even precede the derivation of a model parameter. For example, a narrative WoE was used to determine whether a bioavailability adjustment factor should be used at a dioxin-contaminated site, and then the factor's value was chosen based on a prior quantitative WoE in a published review (Integral Consulting Inc., 2013). Quantitative condition assessments might determine the magnitude and spatial extent of impairment after a site has

been determined impaired by qualitative WoE applied to multiple metrics. In sum, this qualitative WoE determines whether the property to be quantified is real and significant in the context of the assessment.

Table 8-1. Qualities that could be identified by qualitative WoE and the associated quantities that could be derived by the quantitative WoE process (see Figure 3-2)

Example Quality	Example Corresponding Quantity
General cause: teratogen	Sufficient maternal body burden
General cause: carcinogen	Slope factor
Specific cause: chloride in effluent as cause of fish kills	Allowable maximum concentration
Specific cause: ammonia in stream as cause of biological impairment	Total maximum daily load
Bioaccumulative	Bioaccumulation factor
Exposure	Bioavailable concentration at point of contact
Susceptibility	Probability of responding
Alteration	Area of wetland

In some cases, assessments weigh evidence to derive a quantitative result without a prior WoE for a quality. For example, the derivation of ambient water quality criteria for aquatic life is nearly always based on a generic endpoint—protection of aquatic life from effects of direct aqueous exposures on survival, growth, or reproduction. However, for some criteria the generic endpoint is found to not be protective because an important specific effect or route of exposure is not adequately addressed. An example is the water quality criterion for selenium, which addressed a nonstandard effect (skeletal deformities in vertebrates) and route of exposure (exposure in ovo to selenium accumulated from a food web by the female parent). Such examples suggest that it may be desirable to weigh evidence for nonstandard qualities. Even if default qualities are assumed, the first step of WoE, the systematic assembly of evidence, should be performed to obtain all potentially useful quantitative estimates as well as associated information.

The place of these qualitative WoE processes in an overall assessment process depends on the type of assessment. For a conventional risk assessment, a WoE to choose the endpoints (i.e., What effects does the chemical or other agent cause in which potentially exposed organisms?) would be part of the problem formulation. The quantitative analysis and characterization phases follow to quantify the risks. In other cases, separate qualitative and quantitative assessments are performed. For example, a qualitative causal assessment to determine the cause of an observed effect may be followed by a quantitative risk assessment to derive a total maximum daily load, cleanup goal or other quantitative benchmark. In cases like the bioavailability adjustment factor example presented earlier, the qualitative and quantitative WoE processes might both be embedded in the analytical phase of an assessment.

8.2. Weight of Evidence for Deriving the Quantity

If more than one source of data with acceptable relevance and reliability is available to derive a quantitative result, the data sets should be weighed. Conventionally, the data sets would be integrated by combining the quantitative evidence or by choosing the best (i.e., weightiest) evidence. Another approach, the Rule of Five, has been useful in weighing evidence to determine a contaminated site cleanup goal (Box B-2).

8.2.1. Combining quantitative evidence

If the quantitative value can be derived from multiple data sets that are of sufficiently high relevance and reliability, the data sets can be numerically combined (Appendix B). In past EPA practices, this combining has involved taking the arithmetic or geometric mean. For example, multiple acceptable LC_{50} values are available for a species-chemical combination, their geometric mean is used when deriving national ambient water quality criteria (U.S. EPA, 1985). Geometric means of toxicity values are also used to derive soil screening levels (U.S. EPA, 2005a). Rather than treating all evidence equally, the values might be weighted before they are combined. Quantitative weighting could use a quantitative property of the study that influences its reliability such as the inverse variance.

Alternatively, quantities might be weighted based on some qualitative property. In this case, the qualitative weights are converted to numerical equivalents. For example, high, moderate, and low reliability of the study design might be converted to weights of 1, 0.7, and 0.5 or some other values, depending on how much influence study design should have and how great the variance in the study design might be.

Rather than using numerical weights to express the properties of the individual estimates, qualitative weights can be applied to the combined estimate. That is, after a combined value is derived, it might be assigned scores to express one or more qualitative aspects of its relevance or reliability. The weighted mean or other quantity derived from multiple sources should provide greater confidence than a quantity derived from any single source. Otherwise, the evidence should not have been combined.

8.2.2. Choosing the best quantitative evidence

If the range of relevance or reliability of alternative numerical values is large, choosing the best value rather than combining them is advisable. That is, the numerical values should be weighted for relevance and reliability, and the weightiest one used. (Note that in some contexts, established policy might require using the most protective acceptable value to ensure the degree of protection required by law.) Choosing the best value is particularly advisable when values can be derived from multiple types of evidence. For example, if a benchmark value of a contaminant could be derived from laboratory toxicity tests, a mesocosm test or a regional field survey, the results are likely to represent qualitatively different effects and might not be averaged, with or without weighting.

Using WoE to choose the best estimate is illustrated by Table 8-2. This table format presents the magnitude or probability of effects versus integrated weights of evidence for types or groups of evidence as recommended by Hope and Clarkson (2014). The table shows risk assessment results (probability of achieving the effect endpoint) estimated by analysis of each of four evidence groups (numbered circles) with integrated weights derived as in Section 6. In this hypothetical example, Evidence Group 3 might be chosen because it has the highest weight and is not an outlier in terms of the quantitative result (probability of effect). The result might be, Group 3 (evidence derived from stream mesocosm tests) is convincing and gives an estimate of 15% probability of reduced species richness.

Using WoE to choose the best value is not an established practice, but it is recommended for choosing among alternative test data for the same species, endpoint, duration, life stage, and testing conditions by the European Chemical Agency (ECHA, 2008). Although the European Chemical Agency does not specify a method of weighing, it indicates that various studies can provide supporting information to add weight to a particular test.

Table 8-2. WoE matrix to summarize quantitative risk estimates for four evidence types or groups (numbered circles) and their weights. Adapted from Hope and Clarkson (2014).

		Evidence Weight			
		0	+	++	+++
Probability of Achieving Effect Endpoint	>75%				
	51–75%				
	25–50%			①	
	5–24%			②	③
	<5%		④		

8.3. Weighting a Quantitative Result

Quantitative results are often accompanied by statistical expressions of variability or uncertainty, but determining how much weight the result should be given might also be useful. Weighting a quantitative result could inform decision makers and stakeholders about the result's relevance and reliability, irrespective of data scatter (Section 9). For example, an estimate of biological effects might have a small confidence interval, but it could have been generated from laboratory test data with low relevance to the case or might have been generated with poor controls or otherwise unreliable methods. As in weighting qualitative evidence, the specific properties of relevance and reliability listed in Box 5-1 and Box 5-3 and the collective properties listed in Box 6-1 could be included in the weighting of quantitative evidence if they are important.

Strength is generally not a relevant property for weighting quantitative results. For example, an exposure-response model with a high slope is strong, and therefore, lends weight to a causal hypothesis (i.e., a qualitative WoE). However, the result of a quantitative analysis to derive a benchmark exposure has whatever strength it has. An estimate of a benchmark value derived from an exposure-response relationship with a high slope is not given more weight than one derived from a relationship with a low slope.

Weighting might be used to determine the applicability of a number to particular uses. For example, at contaminated sites, weights might be used to distinguish soil benchmark values that are suitable only for screening values from those that might be suitable for cleanup goals. In any case, weights can be used along with statistical uncertainty measures to communicate confidence in results.

9. WEIGHT OF EVIDENCE AND UNCERTAINTY

The process of weighing evidence can be considered a means not only of deriving a conclusion, but also of documenting the assessors' confidence in the conclusion. That is, the weights assigned to pieces and categories of evidence are expressions of how confident the assessors are that the evidence should influence the conclusion in one direction or another. Additionally, the overall weight of the body of evidence is an expression of how much confidence the assessors have in that conclusion. If both laboratory tests and field surveys indicate that a chemical causes spinal deformities in fish, and both categories of evidence are evaluated and scored as highly relevant, strong, and reliable, we are confident that the chemical is a cause of deformities. Even if the WoE analysis is limited to a narrative, the narrative should describe the degree of confidence in the conclusion. For example, the categories of results in the ISAs and IRIS hazard assessments are explicit expressions of confidence in WoE narratives (e.g., "likely to be a causal relationship"—see Table 6-3). Such qualitative expressions of confidence are appropriate for a qualitative conclusion such as the observed relationship is likely causal, the chemical is likely a carcinogen, or the chemical is a contaminant of concern. If the standard scoring system was used to weight evidence, the resulting expression of overall confidence might be that the evidence is convincingly supportive, strongly supportive, or somewhat supportive of the conclusion. The appropriate expression of weight/confidence depends on the assessment and the decision to be supported.

The confidence concerning numerical values should include qualitative weights as well as conventional measures of scatter such as range or confidence limits[Figure 9-1 (Spiegelhalter and Riesch, 2011)].[2] For example, one would have low confidence in a benchmark value that is based on irrelevant evidence, even if it is statistically precise. Hence, presentations of quantitative results might have four parts:

1. The quality expressed—threshold for reproductive effects
2. The numerical result and units—*x* mg/L.
3. Scatter—95% CI of ±0.2*x* mg/L.
4. Weight—the body of evidence is highly relevant and moderately reliable, diverse, and coherent.

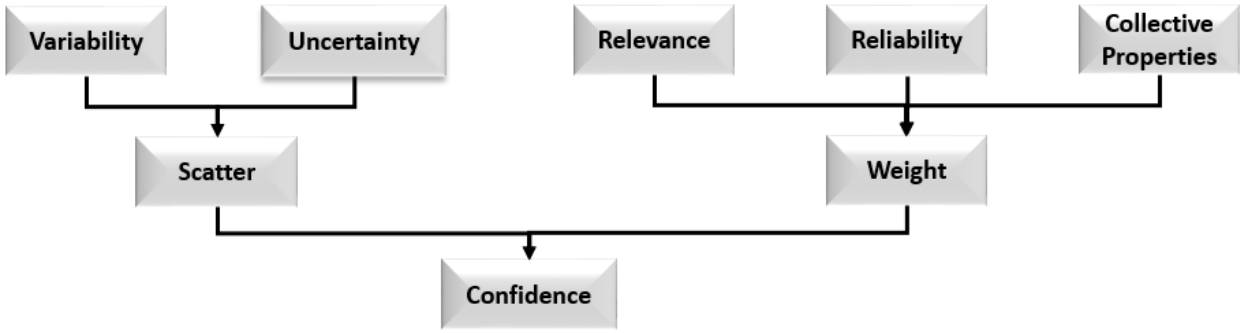


Figure 9-1. A diagram of the combination of statistical scatter and qualitative weight to define the confidence that should be afforded an assessment result.

[2] This discussion assumes an objectivist view that conceives uncertainty and variability as probabilities based on frequencies, which are distinct from qualitative judgments of weight. A subjectivist would consider all components of Figure 9-1 to be encompassed by degree of belief, and all components could be expressed as subjective probabilities.

The pieces and types of evidence that are not used to derive the numerical result could still provide information concerning the result. For example, another piece of evidence might place a limit on the possible range of effects and that information can be used to truncate the confidence limit on the estimated effect. Other evidence might inform the scope of a result. For example, a field-based benchmark value for a contaminant in streams might be based on benthic invertebrates because they provide the best data. Fish survey data that are not sufficient to derive a benchmark, however, might be sufficient to conclude that the invertebrate-based benchmark is likely also protective of the fish assemblage.

Conceivably, assessors could address the uncertainty concerning the confidence expressed by the WoE. No methods have been found for estimating or describing such meta-uncertainties. An estimate of WoE uncertainty, however, can be obtained by replicating the assessment. That is, multiple assessment teams could be engaged to assemble, weight, and weigh the evidence so that the variance among WoE results might be estimated. That implies an extraordinary effort for a routine assessment, but replication of a WoE for a condition assessment has been done experimentally (Bay et al., 2007). Six experts were provided conventional sediment quality triad data for 25 embayment sites in California. They were asked to rank the sites from best to worst and to categorize them into one of six standard categories. No instructions were provided, so the participants apparently used whatever WoE method they used in their professional practice. The rankings were highly correlated (mean Spearman rank correlation = 0.92). Differences in categorization of a site were common but mostly small and were primarily due to differences in relative weighting of the three types of evidence, so common weighting guidelines might have reduced discrepancies.

A sensitivity analysis would be more practical than replicating the assessment as a means of determining the variability in the weighting process. That is, assessors could determine the influence of changing the weights on WoE results. If the choice between alternative hypotheses could change when, for example, the reliability of a piece of evidence was scored as + rather than ++, that sensitivity would be noted. Such cases would tend to occur when few pieces of evidence are considered, when the overall weight for a body of evidence is marginal, or the bodies of evidence for alternative hypotheses have similar weights.

10. WEIGHT-OF-EVIDENCE SUMMARY AND THE PATH FORWARD

The basic framework of assembling evidence, weighting it, and weighing the body of evidence is fundamental to making inferences based on WoE (Figure 10-1). Even if a WoE is not explicitly performed using these steps, assessors should at least consider each of these processes. Adhering to an explicit framework can improve the WoE results, enhance transparency, and increase the confidence of reviewers, stakeholders, and decision makers.

The steps to implement the framework for qualitative WoE, with section numbers, include:

- Finding information (4.2.1),
- Design and conduct studies to fill information gaps (4.3)
- Screening studies (4.2.2),
- Categorizing studies (4.2.3),
- Deriving evidence from the data (4.2.4),
- Evaluating the evidence with respect to properties (5.3),
- Assigning weight scores based on the evaluation (5.3),
- Creating a scoring table to summarize the results of evidence weighting (5.4),
- Integrating the weighted evidence into bodies of evidence for each hypothesis (6.2),
- Creating a WoE table to summarize the results of evidence integration (6.2),
- Interpreting the bodies of evidence (6.3),
- Explaining ambiguities and discrepancies (6.4),
- Reiterating the process if necessary (6.4), and
- Presenting results (6.5).

In addition to the approach for applying WoE to qualities such as causality and impairment, this document briefly addresses WoE for quantities and how WoE for qualities and quantities are related (Section 8). Although WoE for quantities such as benchmark or parameter values is uncommon now, it seems likely that the use of meta-analysis and other techniques for weighting and combining quantities will gain increasing use. A combined approach to WoE for qualities and quantities is likely advantageous.

Similarly, this document briefly discusses the relationship between uncertainty and WoE (Section 9). By combining the somewhat narrow concept of uncertainty (i.e., measures of the scatter of data or of estimates) with the broader concept that some evidence and conclusions deserve more weight, it should be possible to improve the communication of confidence in assessment results.

The WoE approach in this document is based on experience in the EPA, particularly with determining the causes of impairment in ecosystems (http://www3.epa.gov/caddis/). Its application in practice will vary among programs due to differences in issues, policies, prior practices, and the amount and variety of evidence potentially available for weighing. As implementation of this approach progresses, context-specific guidance is expected to be developed in the form of exemplary applications of WoE or program-specific WoE guidance documents.

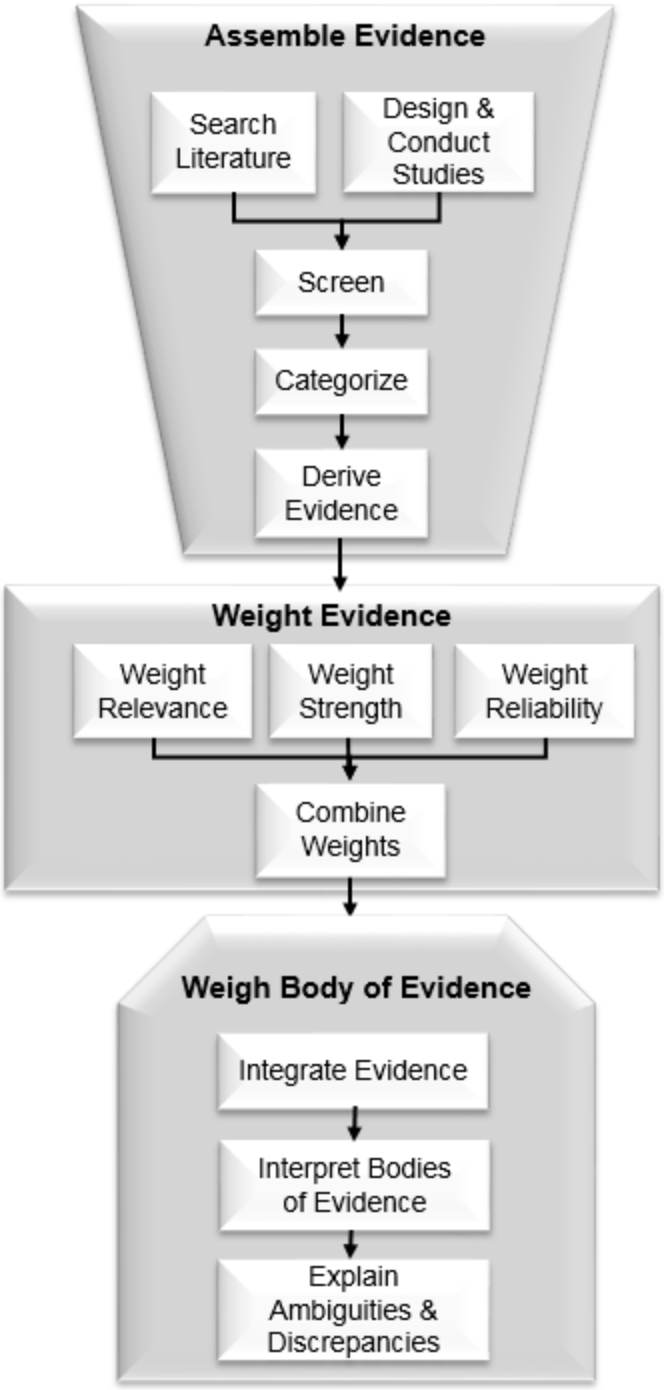


Figure 10-1. The detailed framework for qualitative WoE.

In summary, this document is intended to help ecological assessors improve the practice of WoE without imposing burdensome prescriptions. If read and applied in that spirit, this document should help advance the cause of well-informed environmental protection.

11. REFERENCES

Anchor, Q. (2013). Baseline-ecological risk assessment report, Patrick Bayou Superfund Site, Deer Park, Texas. Ocean Springs, MS: Patrick Bayou Joint Defense Group and U.S. EPA.

Anderson, D. (2008). Model based inference in the life sciences. New York, NY: Springer Science and Business Media.

Ankley, GT; Bennett, RS; Erickson, RJ; Hoff, DJ; Hornung, MW; Johnson, RD; Mount, DR; Nichols, JW; Russom, CL; Schmieder, PK; Serrrano, JA; Tietge, JE; Villeneuve, DL. (2010). Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ Toxicol Chem 29: 730-741. http://www.ncbi.nlm.nih.gov/pubmed/20821501

Batley, GE; Burton, GA; Chapman, PM; Forbes, VE. (2002). Uncertainties in sediment quality weight-of-evidence (WOE) assessments. Hum Ecol Risk Assess 8: 1517-1547. http://dx.doi.org/10.1080/20028091057466

Bay, S; Berry, W; Chapman, PM; Fairey, R; Gries, T; Long, E; MacDonald, D; Weisberg, SB. (2007). Evaluating consistency of best professional judgment in the application of a multiple lines of evidence sediment quality triad. Integr Environ Assess Manag 3: 491-497. http://www.ncbi.nlm.nih.gov/pubmed/18046798

Becker, RA; Ankley, GT; Edwards, SW; Kennedy, SW; Linkov, I; Meek, B; Sachana, M; Segner, H; Van Der Burg, B; Villeneuve, DL; Watanabe, H; Barton-Maclaren, TS. (2015). Increasing scientific confidence in adverse outcome pathways: Application of tailored Bradford-Hill considerations for evaluating weight of evidence. Regul Toxicol Pharmacol 72: 514-537. http://www.ncbi.nlm.nih.gov/pubmed/25863193

Benedetti, M; Ciaprini, F; Piva, F; Onorati, F; Fattorini, D; Notti, A; Ausili, A; Regoli, F. (2012). A multidisciplinary weight of evidence approach for classifying polluted sediments: Integrating sediment chemistry, bioavailability, biomarkers responses and bioassays. Environ Int 38: 17-28. http://www.ncbi.nlm.nih.gov/pubmed/21982029

Bilotta, G; Milner, A; Boyd, I. (2014). On the use of systematic reviews to inform environmental policies. Environ Sci Pol 42: 67-77.

Bilyard, G; Beckert, H; Bascietto, J; Abrams, C; Dyer, S; Haselow, L. (1997). Using the data quality objectives process during the design and conduct of ecological risk assessment. Washington, DC: U.S. Department of Energy, Office of Environmental Policy and Assistance. http://www.monitor2manage.com.au/userdata/downloads/p_/Risk%20management%20and%20DQO.pdf

Black & Veatch Special Projects Corps. (2011). Ecological risk assessment for the estuary at the LCP chemical site in Brunswick, Georgia: Site investigation/analysis and risk characterization (Revision 4). Atlanta, GA: U.S. Environmental Protection Agency, Region 4. http://www.epa.gov/sites/production/files/2014-03/documents/baseline_ecological_risk_assessment_april2011pdf.pdf

Blocksom, K; Johnson, B. (2009). Development of a regional macroinvertebrate index for large river bioassessments. Ecol Appl 9: 313-328.

Blyth, CR. (1972). On Simpson's paradox and the sure-thing principle. J Am Stat Assoc 67: 364-366. http://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10482387

Bombardier, M; Bermingham, N. (1999). The SED-TOX index: toxicity-directed management tool to assess and rank sediments based on their hazard - concept and application. Environ Toxicol Chem 18: 685-698.

Bombardier, M; Blaise, C. (2000). Comparative study of the sediment-toxicity index, benthic community metrics and contaminant concentrations. Water Qual Res J Can 35: 753-780.

Borenstein, M; Hedges, L; Higgins, J; Rothstein, H. (2009). Introduction to meta-analysis. Chichester, U.K.: Wiley.

Carriger, J; Barron, M. (2016). A practical probabilistic graphical modeling tool for weighing ecological risk-based evidence. Soil Sed Contam 25: 476-487.

Carriger, JF; Barron, MG. (2011). Minimizing risks from spilled oil to ecosystem services using influence diagrams: The deepwater horizon spill response. Environ Sci Technol 45: 7631-7639. http://dx.doi.org/10.1021/es201037u

CEE (Collaboration for Environmental Evidence). (2013). Guidelines for systematic reviews in environmental management. Version 4.2. Banker, UK: CEE.

Chapman, P. (1990). The sediment quality triad approach to determining pollution-induced degradation. Sci Total Environ 97/98: 815-825.

Chapman, PM. (1996). Presentation and interpretation of Sediment Quality Triad data. Ecotoxicology 5: 327-339. http://www.ncbi.nlm.nih.gov/pubmed/24193872

Chapman, PM. (2007). Determining when contamination is pollution - weight of evidence determinations for sediments and effluents. Environ Int 33: 492-501. http://www.ncbi.nlm.nih.gov/pubmed/17027966

Chapman, PM; Anderson, J. (2005). A decision-making framework for sediment contamination. Integr Environ Assess Manag 1: 163-173. http://www.ncbi.nlm.nih.gov/pubmed/16639882

COA Sediment Task Group. (2008). Canada-Ontario decision-making framework for assessment of Great Lakes contaminated sediments. Environment Canada and Ontario Ministry of the Environment. http://publications.gc.ca/collections/collection_2010/ec/En164-14-2007-eng.pdf

Collier, ZA; Gust, KA; Gonzalez-Morales, B; Gong, P; Wilbanks, MS; Linkov, I; Perkins, EJ. (2016). A weight of evidence assessment approach for adverse outcome pathways. Regul Toxicol Pharmacol 75: 46-57. http://www.ncbi.nlm.nih.gov/pubmed/26724267

Cormier, SM; Suter, GW, 2nd. (2008). A framework for fully integrating environmental assessment. Environ Manage 42: 543-556. http://www.ncbi.nlm.nih.gov/pubmed/18506517

Cormier, SM; Suter, GW, 2nd. (2013). A method for assessing causation of field exposure-response relationships. Environ Toxicol Chem 32: 272-276. http://www.ncbi.nlm.nih.gov/pubmed/23161561

Cormier, SM; Suter, GW, 2nd; Zheng, L; Pond, GJ. (2013). Assessing causation of the extirpation of stream macroinvertebrates by a mixture of ions. Environ Toxicol Chem 32: 277-287. http://www.ncbi.nlm.nih.gov/pubmed/23147750

Cormier, SM; Suter, GW; Norton, SB. (2010). Causal characteristics for ecoepidemiology. Hum Ecol Risk Assess 16: 53-73. http://dx.doi.org/10.1080/10807030903459320

CRD (Centre for Reviews and Dissemination). (2009). Systematic reviews: CRD's guidance for undertaking reviews in health care. York, U.K.: Centre for Reviews and Dissemination, University of York.

Dagnino, A; Sforzini, S; Dondero, F; Fenoglio, S; Bona, E; Jensen, J; Viarengo, A. (2008). A "weight of evidence" approach for the integration of environmental "triad" data to assess ecological risk and biological vulnerability. Integr Environ Assess Manag 4: 314-326. http://www.ncbi.nlm.nih.gov/pubmed/18393577

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using –point, 7-point and 10-point scales. Int J Market Res 50: 61-77.

Doi, SA; Thalib, L. (2008). A quality-effects model for meta-analysis. Epidemiology 19: 94-100. http://www.ncbi.nlm.nih.gov/pubmed/18090860

Douglas, H. (2012). Weighing complex evidence in a democratic society. Kennedy Inst Ethics J 22: 139-162. http://www.ncbi.nlm.nih.gov/pubmed/23002581

Duboudin, C; Ciffroy, P; Magaud, H. (2004). Effects of data manipulation and statistical methods on species sensitivity distributions. Environ Toxicol Chem 23: 489-499. http://www.ncbi.nlm.nih.gov/pubmed/14982398

EC (European Commission). (2013). Introduction to the new EU water framework directive. http://ec.europa.eu/environment/water/water-framework/info/intro_en.htm.

ECETOC (European Centre for Ecotoxicology and Toxicology of Chemicals). (2009). Framework for the integration of human and animal data in chemical risk assessment. (Technical Report No. 104).

Brussels, Belgium: ECETOC. http://www.ecetoc.org/uploads/Publications/documents/TR%20104.pdf
ECHA (European Chemicals Agency). (2008). Guidance on information requirements and chemical safety assessment. Chapter R.10: characterization of dose [concentration]-response for environment. Heksinki, Finland: ECHA. http://echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment
ECHA. (2010). Practical guide 2: How to report weight of evidence. (ECHA-10-B-05-EN). Helsinki, Finland: ECHA. http://echa.europa.eu/documents/10162/13655/pg_report_weight_of_evidence_en.pdf
ECHA. (2015). Read-across assessment framework (RAAF). (ECHA-15-R-07-EN). Helsinki, Finland: ECHA. http://echa.europa.eu/documents/10162/13628/raaf_en.pdf
Egger, M; Juni, P; Bartlett, C; Holenstein, F; Sterne, J. (2003). How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 7: 1-76. http://www.ncbi.nlm.nih.gov/pubmed/12583822
Fenton, N; Neil, M. (2013). Risk Assessment and decision analysis with Bayesian Networks. Bocan Raton, Fl: CRC Press.
Ferguson, CJ; Brannick, MT. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. Psychol Methods 17: 120-128. http://www.ncbi.nlm.nih.gov/pubmed/21787082
Forbes, VE; Calow, P. (2002). Species sensitivity distributions revisited: A critical appraisal. Hum Ecol Risk Assess 8: 473-492. http://www.tandfonline.com/doi/abs/10.1080/10807030290879781
Fox, GA. (1991). Practical causal inference for ecoepidemiologists. J Toxicol Environ Health 33: 359-373. http://www.ncbi.nlm.nih.gov/pubmed/1875428
Good, I. (1950). Probability and the weighing of evidence. New York, NY: Hafner Press.
Gough, D; Oliver, S; Thomas, J. (2012). An introduction to systematic reviews. London, U.K.: Sage Publications.
Grapentine, L; Anderson, J; Boyd, D; Burton, GA; DeBarros, C; Johnson, G; Marvin, C; Milani, D; Painter, S; Pascoe, T; Reynoldson, T; Richman, L; Solomon, K; Chapman, PM. (2002). A decision making framework for sediment assessment developed for the Great Lakes. Hum Ecol Risk Assess 8: 1641-1655. http://dx.doi.org/10.1080/20028091057538
Gray, G. (1994). Complete risk characterization. In Risk in Perspective (pp. 1-2). Harvard Center for Risk Analysis. https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1273/2013/06/Complete-Risk-Characterization-Nov-1994.pdf
Greenberg, M; Charters, D. (2007). The rule of five: A novel approach to derive PRGs. Joint Services Environmental Management Conference, May 21-24, 2007, Columbus, OH.
Greenland, S; O'Rourke, K. (2001). On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2: 463-471. http://www.ncbi.nlm.nih.gov/pubmed/12933636
Hawkins, CP. (2006). Quantifying biological integrity by taxonomic completeness: its utility in regional and global assessments. Ecol Appl 16: 1277-1294. http://www.ncbi.nlm.nih.gov/pubmed/16937797
Hertzberg, RC; Teuschler, LK. (2002). Evaluating quantitative formulas for dose-response assessment of chemical mixtures. Environ Health Perspect 110 Suppl 6: 965-970. http://www.ncbi.nlm.nih.gov/pubmed/12634126
Higgins, J; Green, S. (2011). Cochrane handbook for systematic reviews of interventions. Version 5.1.0. Cambridge, U.K.: The Cochrane Collaboration. http://handbook.cochrane.org/
Hilborn, R; Mangel, M. (1997). The ecological detective: Confronting models with data. Monographs in population biology. Princeton, NJ: Prince University Press.
Hill, AB. (1965). The environment and disease: association or causation? Proc R Soc Med 58: 295-300. http://www.ncbi.nlm.nih.gov/pubmed/14283879

Hope, BK; Clarkson, JR. (2014). A strategy for using weight-of-evidence methods in ecological risk assessments. Hum Ecol Risk Assess 20: 290-315. http://dx.doi.org/10.1080/10807039.2013.781849

Hume, D. (1748). An enquiry concerning human understanding. Amherst, NY: Prometheus.

Hunter, JE; Schmidt, FL. (2004). Methods of meta-analysis: Correcting error and bias in research findings (Second ed.). Thousand Oaks, CA: Sage Publications.

Integral Consulting Inc. (2013). Baseline ecological risk assessment: San Jacinto River Waste Pits Superfund Site. (9559469). Seattle, WA. http://semspub.epa.gov/src/collection/06/SC32405

Jaworska, J; Gabbert, S; Aldenberg, T. (2010). Towards optimization of chemical testing under REACH: a Bayesian network approach to Integrated Testing Strategies. Regul Toxicol Pharmacol 57: 157-167. http://www.ncbi.nlm.nih.gov/pubmed/20156511

Johnston, RK; Munns, WR, Jr.; Tyler, PL; Marajh-Whittemore, P; Finkelstein, K; Munney, K; Short, FT; Melville, A; Hahn, SP. (2002). Weighing the evidence of ecological risk from chemical contamination in the estuarine environment adjacent to the Portsmouth Naval Shipyard, Kittery, Maine, USA. Environ Toxicol Chem 21: 182-194. http://www.ncbi.nlm.nih.gov/pubmed/11804053

Karr, JR; Fausch, KD; Angermeier, PL; Yant, PR; Schlosser, IJ. (1986). Assessing biological integrity in running waters: A method and its rationale. (Publication 5). Champaign, Illinois: Illinois Natural History Survey Special. http://www.nrem.iastate.edu/class/assets/aecl518/Discussion%20Readings/Karr_et_al._1986.pdf

Keilser, J; Collier, Z; Chu, E; Sinatra, N; Linkov, L. (2014). Value of information analysis: the state of application. Environ Syst Decis 34: 3-23.

Krimsky, S. (2005). The weight of scientific evidence in policy and law. Am J Public Health 95 Suppl 1: S129-136. http://www.ncbi.nlm.nih.gov/pubmed/16030328

Laplace, P. (1812). A philosophical essay on probabilities. 1902 translation. New York, NY: John Wiley.

Linkov, I; Loney, D; Cormier, S; Satterstrom, FK; Bridges, T. (2009). Weight-of-evidence evaluation in environmental assessment: review of qualitative and quantitative approaches. Sci Total Environ 407: 5199-5205. http://www.ncbi.nlm.nih.gov/pubmed/19619890

Linkov, I; Massey, O; Keisler, J; Rusyn, I; Hartung, T. (2015). From "weight of evidence" to quantitative data integration using multicriteria decision analysis and Bayesian methods. ALTEX 32: 3-8. http://www.ncbi.nlm.nih.gov/pubmed/25592482

Linkov, I; Moberg, E. (2012). Multi-criteria decision analysis: Environmental applications and case studies. Boca Raton, FL: CRC Press.

Luftig, SD. (1999). Issuance of final guidance: Ecological risk assessment and risk management principles for superfund sites. Memorandum, October 7. (OSWER Directive 9285.7-28 P). Washington, D.C.: Office of Emergency and Remedial Response, U.S. EPA. http://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=9100L92P.TXT

McDonald, BG; deBruyn, AM; Wernick, BG; Patterson, L; Pellerin, N; Chapman, PM. (2007). Design and application of a transparent and scalable weight-of-evidence framework: an example from Wabamun Lake, Alberta, Canada. Integr Environ Assess Manag 3: 476-483. http://www.ncbi.nlm.nih.gov/pubmed/18046796

McGarity, T; Wagner, W. (2008). Bending science: How special interests corrupt public health research. Cambridge, MA: Harvard University Press.

McPherson, C; Chapman, PM; Debruyn, AM; Cooper, L. (2008). The importance of benthos in weight of evidence sediment assessments--a case study. Sci Total Environ 394: 252-264. http://www.ncbi.nlm.nih.gov/pubmed/18295824

Meek, ME; Palermo, CM; Bachman, AN; North, CM; Jeffrey Lewis, R. (2014). Mode of action human relevance (species concordance) framework: Evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. J Appl Toxicol 34: 595-606. http://www.ncbi.nlm.nih.gov/pubmed/24777878

Menzie, C; Henning, MH; Cura, J; Finkelstein, K; Gentile, J; Maughan, J; Mitchell, D; Petron, S; Potocki, B; Svirsky, S; Tyler, P. (1996). Special report of the Massachusetts weight-of-evidence workgroup A weight-of-evidence approach for evaluating ecological risks. Hum Ecol Risk Assess 2: 277-304. http://dx.doi.org/10.1080/10807039609383609

Murphy, BL; Morrison, RD. (2002). Introduction to environmental forensics. San Diego, CA: Academic Press.

Newman, MC; Clements, WH. (2008). Ecotoxicology: A comprehensive treatment. Boca Raton, FL: CRC Press.

Norton, SB; Cormier, SM; Suter, GW. (2014). Ecological causal assessment. Boca Raton, FL: CRC Press.

NRC (National Research Council). (1983). Risk assessment in the Federal Government: managing the process. Washington, D.C.: The National Academies Press. http://www.nap.edu/openbook.php?isbn=0309033497

NRC. (2014). Review of EPA's integrated risk information system (IRIS) process. Washington, D.C.: The National Academies Press. http://www.nap.edu/catalog/18764/review-of-epas-integrated-risk-information-system-iris-process

OECD (Organization for Economic Cooperation and Development). (2013). Guidance document on developing and assessing adverse outcome pathways. Series on testing and assessment No. 184. Paris, France: Organization for Economic Cooperation and Deelopment. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2013)6&doclanguage=en

Pond, G; Passmore, M; Borsuk, F; Reynolds, L; Rose, C. (2008). Downstream Effects of Mountaintop Coal Mining: Comparing Biological Conditions using Family- and Genus-Level Macroinvertebrate Bioassessment Tools. 27: 717-737.

Pope, C; Mays, N; Popay, J. (2007). Synthesizing qualitative and quantitative health evidence: A guide to methods. Maidenhead, U.K.: Open University Press.

Ralof, J. (2010). Atrazine paper's challenge: Who's resposible for accuracy. Science News, May 6, 2010.

Rhomberg, L. (2015). Hypothesis-based weight of evidence: An approach to assessing causation and its application to regulatory toxicology. Risk Anal 35: 1114-1124. http://www.ncbi.nlm.nih.gov/pubmed/24724710

Rhomberg, LR; Goodman, JE; Bailey, LA; Prueitt, RL; Beck, NB; Bevan, C; Honeycutt, M; Kaminski, NE; Paoli, G; Pottenger, LH; Scherer, RW; Wise, KC; Becker, RA. (2013). A survey of frameworks for best practices in weight-of-evidence analyses. Crit Rev Toxicol 43: 753-784. http://www.ncbi.nlm.nih.gov/pubmed/24040995

Rooney, AA; Boyles, AL; Wolfe, MS; Bucher, JR; Thayer, KA. (2014). Systematic review and evidence integration for literature-based environmental health science assessments. Environ Health Perspect 122: 711-718. http://www.ncbi.nlm.nih.gov/pubmed/24755067

SAB (Scientific Advisory Board). (2012). SAB review of the EPA's ecological assessment action plan. (EPA-SAB-12-010). Washington, D.C.: SAB, U.S. EPA. http://yosemite.epa.gov/sab/sabproduct.nsf/773C41AF81B7B16C85257A8700796DA9/$File/EPA-SAB-12-010-unsigned.pdf

SAB. (2014). SAB review of the draft EPA report *Connectivity of Streams and Wetlands to Downstream Waters: A Review and Synthesis of the Scientific Evidence*. Letter to Administrator Gina McCarthy. (EPA-SAB-15-001). Washington, D.C.: SAB, U.S. EPA. http://yosemite.epa.gov/sab/sabproduct.nsf/WebBoard/AF1A28537854F8AB85257D74005003D2/$File/EPA-SAB-15-001+unsigned.pdf

SAB. (2015). SAB review of the EPA's *Evaluation of the Inhalation Carcinogenicity of Ethylene Oxide (Revised External Review Draft - August 2014).* Letter to Administrator Gina McCarthy. (EPA-SAB-15-002). Washington, D.C.: SAB, U.S. EPA. http://yosemite.epa.gov/sab/sabproduct.nsf/WebBoard/BD2B2DB4F84146A585257E9A0070E655/$File/EPA-SAB-15-012+unsigned.pdf

Sanz-Martin, M; Pitt, K; Condon, R; Lucus, C; de Santana, C; Duarte, C. (2016). Flawed citation practices facilitate the unsubstantiated perception of a global trend toward increased jellyfish blooms. Glob Ecol Biogeo 25: 1039-1049.
Semenzin, E; Critto, A; Rutgers, M; Marcomini, A. (2008). Integration of bioavailability, ecology and ecotoxicology by three lines of evidence into ecological risk indexes for contaminated soil assessment. Sci Total Environ 389: 71-86. http://www.ncbi.nlm.nih.gov/pubmed/17904618
Semenzin, E; Critto, A; Rutgers, M; Marcomini, A. (2009). DSS-ERAMANIA: Decision support system for site-specific ecological risk assessment of contaminated sites. In A Marcomini; GWI Suter; A Critto (Eds.), Decision Support Systems for Risk-Based Management of Contaminated Sites. New York, NY: Springer.
Semenzin, E; Lanzellotto, E; Hristozov, D; Critto, A; Zabeo, A; Giubilato, E; Marcomini, A. (2015). Species sensitivity weighted distribution for ecological risk assessment of engineered nanomaterials: the n-TiO2 case study. Environ Toxicol Chem 34: 2644-2659. http://www.ncbi.nlm.nih.gov/pubmed/26058704
Shull, D; Pulket, M. (2015). Causal analysis of the smallmouth bass decline in the Susquehanna and Juniata Rivers. Harrisburg, PA: Pennsylvania Department of Environmental Protection. http://files.dep.state.pa.us/Water/Drinking%20Water%20and%20Facility%20Regulation/WaterQualityPortalFiles/SusquehannaRiverStudyUpdates/SMB_CADDIS_Report.pdf
Small, MJ. (2008). Methods for assessing uncertainty in fundamental assumptions and associated models for cancer risk assessment. Risk Anal 28: 1289-1308. http://www.ncbi.nlm.nih.gov/pubmed/18844862
Smith, E; Lipkovich, I; Ye, K. (2002). Weight-of-Evidence (WOE): Quantitative estimation of probability of impairment for individual and multiple lines of evidence. Hum Ecol Risk Assess 8: 1585-1596.
Spiegelhalter, D; Riesch, H. (2011). Don't know, can't know: Embracing deeper uncertainties when analyzing risks. Phil Trans Royal Soc 369: 4730-4750.
Stahl, C; Cimorelli, A; Chow, A. (2002). A new approach to environmental decision analysis: Multi-criteria integrated resource assessment (MIRA). Bull Sci Technol Soc 22: 443-459.
Susser, M. (1986). Rules of inference in epidemiology. Regul Toxicol Pharmacol 6: 116-128. http://www.ncbi.nlm.nih.gov/pubmed/2941827
Suter, GW, 2nd; Cormier, S. (2011). Why and how to combine evidence in environmental assessments: Weighing evidence and building cases. Sci Total Environ 409: 1406-1417.
Suter, GW, 2nd; Cormier, S. (2014). The problem of biased data and potential solutions for health and environmental assessments. Hum Ecol Risk Assess 21: 1736-1752.
Suter, GW, II. (1993). Ecological risk assessment. Boca Raton, FL: Lewis Publishers.
Suter, GW, II. (1996). Risk characterization for ecological risk assessment of contaminated sites. (ES/ER/TM-200). Oak Ridge, TN: Oak Ridge National Laboratory. http://rais.ornl.gov/documents/tm200.pdf
Suter, GW, II; Efroymson, RA; Sample, BE; Jones, DS. (2000). Ecological risk assessment for contaminated sites. Boca Raton, FL: Lewis Publishers.
Suter, GW, II; Traas, T; Posthuma, L. (2002). Issues and practices in the derivation and use of species sensitivity distributions. In L Posthuma; GW Suter, II; T Traas (Eds.), Species Sensitivity Distributions in Ecotoxicology (pp. 437-474). Boca Raton, FL: Lewis Publishers.
Todd, P; Yeo, D; Li, D; Ladle, R. (2007). Citing practices in ecology: can we believe our own words? Oikos 116: 1599-1601.
Tood, P; Guest, J; Lu, J; Chou, L. (2010). One in four citations in marine biology papers is inappropriate. Mar Ecol Prog Ser 408: 299-303.
Turner, R; Speiegelhalter, D; Smith, G; Thompson, S. (2009). Bias modeling in evidence synthesis. J R Stat Soc 172: 21-47.
U.S. EPA (Environmental Protection Agency). (1985). Guidelines for deriving numeric national water quality criteria for the protection of aquatic organisms and their uses. (PB85-227049). Washington, D.C.: U.S. EPA. http://www.epa.gov/sites/production/files/2015-

08/documents/guidelines_for_deriving_nnwqc_for_the_protectin_of_aquatic_organisms_and_their_uses.pdf
U.S. EPA (Environmental Protection Agency). (1992a). Guidance for data useability in risk assessment (Part A). (EPA/540/R-92/003). Washington, DC: Office of Emergency and Remedial Response. https://rais.ornl.gov/documents/USERISKA.pdf
U.S. EPA (Environmental Protection Agency). (1992b). Guidance for data useability in risk assessment (part B). (9285.7-09B, PB92 -963362). Washington, DC: Office of Emergency and Remedial Response, U.S. EPA. http://rais.ornl.gov/documents/USERISKB.pdf
U.S. EPA (Environmental Protection Agency). (1994). Guidance for the data quality objectives process: EPA QA/G-4. (EPA/600/R-96/055). Washington, D.C.: Office of Research and Development, U.S. EPA. http://www3.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/qa/epaqag4.pdf
U.S. EPA (Environmental Protection Agency). (1996). Biological criteria: Technical guidance for streams and small rivers. (EPA/822/B-096/001). Washington, D.C.: Office of Water, U.S. EPA. http://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=20003GSJ.TXT
U.S. EPA (Environmental Protection Agency). (1998). Guidelines for ecological risk assessment. (EPA/630/R-95/002F). Washington, D.C.: U.S. EPA. http://www.epa.gov/raf/publications/pdfs/ecotxtbx.pdf
U.S. EPA (Environmental Protection Agency). (2000). Stressor identification guidance document. (EPA/822/B-00/025). Washington, D.C.: Office of Water, Office of Research and Development, U.S. EPA. http://permanent.access.gpo.gov/websites/epagov/www.epa.gov/ost/biocriteria/stressors/stressorid.pdf
U.S. EPA (Environmental Protection Agency). (2002a). Guidance for quality assurance project plans: EPA QA/G-5. (EPA/240/R-02/009). Washington, D.C.: Office of Environmental Information, U.S. EPA. http://www.epa.gov/quality/qs-docs/g5-final.pdf
U.S. EPA (Environmental Protection Agency). (2002b). Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by the Environmental Protection Agency. (EPA/260/R-02/008). Washington, D.C.: Office of Environmental Information, U.S. EPA. http://www.epa.gov/quality/informationguidelines/documents/EPA_InfoQualityGuidelines.pdf
U.S. EPA (Environmental Protection Agency). (2005a). Guidance for developing ecological soil screening levels (ECO-SSLs). (Publication 9285.7-55). Washington, D.C.: Office of Solid Waste and Emergency Response, U.S. EPA. http://rais.ornl.gov/documents/ecossl.pdf
U.S. EPA (Environmental Protection Agency). (2005b). Guidelines for carcinogen risk assessment (pp. 166). (EPA/630/P-03/001F). Washington, D.C.: Risk Assessment Forum, Office of Research and Development, U.S. EPA. http://www2.epa.gov/osa/guidelines-carcinogen-risk-assessment
U.S. EPA (Environmental Protection Agency). (2008). EPA quality policy. (EPA Order CIO 2106.0). Washington, D.C.: Office of Environmental Information, U.S. EPA. http://www.epa.gov/sites/production/files/2015-09/documents/epa_order_cio_21060.pdf
U.S. EPA (Environmental Protection Agency). (2009a). Analysis of the causes of a decline in the san joaquin kit fox population on the Elk Hills, Naval Petroleum Reserve #1, California. (EPA/600/R-08/130). Cincinnati, OH: National Center for Environmental Assessment, Office of Research and Development, U.S. EPA. http://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=200367&CFID=53016685&CFTOKEN=36726798
U.S. EPA (Environmental Protection Agency). (2009b). Guidance on the development, evaluation, and application of environmental models. (EPA/100/K-09/003). Washington, D.C.: Council for Regulatory Environmental Modeling, Office of the Science Advisor, U.S. EPA. http://www.epa.gov/crem/library/cred_guidance_0309.pdf

U.S. EPA (Environmental Protection Agency). (2010a). Inferring causes of biological impairment in the Clear Fork Watershed, West Virginia. (EPA/600/R-08/146). Washington, D.C.: National Center for Environmental Assessment, Office of Research and Development, U.S. EPA. http://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=201963&CFID=61073289&CFTOKEN=23932460

U.S. EPA (Environmental Protection Agency). (2010b). Integrating ecological assessment and decision-making at EPA: a path forward. Results of a colloquium in response to Science Advisory Board and National Research Council recommendations. (EPAS/100/R-10/004). Washington, D.C.: Risk Assessment Forum, Office of Research and Development, U.S. EPA. http://www.epa.gov/sites/production/files/2013-09/documents/integrating-ecolog-assess-decision-making.pdf

U.S. EPA (Environmental Protection Agency). (2011a). Evaluation guidelines for ecological toxicity data in the open literature. Washington, D.C.: Office of Pesticide Programs, U.S. EPA. http://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/evaluation-guidelines-ecological-toxicity-data-open

U.S. EPA (Environmental Protection Agency). (2011b). A field-based aquatic life benchmark for conductivity in Central Appalachian Streams. (EPA/600/R-10/023F). Washington, DC: Office of Research and Development, National Center for Environmental Assessment, U.S. EPA. http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=233809

U.S. EPA (Environmental Protection Agency). (2011c). Stressor identification (SI) at contaminated sites: Upper Arkansas River, Colorado (EPA/600/R-08/029). Washington, D.C.: National Center for Environmental Assessment, Office of Research and Development, U.S. EPA. http://cfpub.epa.gov/ncea/caddis/recordisplay.cfm?deid=189290

U.S. EPA (Environmental Protection Agency). (2012a). Benchmark dose technical guidance. (EPA/100/R-12/001). Washington, D.C.: Risk Assessment Forum, U.S. EPA. http://www2.epa.gov/osa/benchmark-dose-technical-guidance

U.S. EPA (Environmental Protection Agency). (2012b). Weight-of-evidence: evaluating results of EDSP Tier 1 screening to identify the need for Tier 2 testing (pp. 47). Washington, D.C.: Office of Chemical Safety and Pollution Prevention, U.S. EPA. http://www.regulations.gov/#!documentDetail;D=EPA-HQ-OPPT-2010-0877-0021

U.S. EPA (Environmental Protection Agency). (2013). Integrated science assessment for lead. (EPA/600/R-10/075F). Research Triangle Park, NC: National Center for Environmental Assessment, Office of Research and Development, U.S. EPA. http://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=255721

U.S. EPA (Environmental Protection Agency). (2014a). An assessment of potential mining impacts on salmon ecosystems of Bristol Bay, Alaska. Seattle, WA: Region 10, U.S. EPA. http://cfpub.epa.gov/ncea/bristolbay/recordisplay.cfm?deid=253500

U.S. EPA (Environmental Protection Agency). (2014b). Welfare risk and exposure assessment for ozone. (EPA/452/R-14/005a). Washington, D.C.: Office of Air Quality Planning and Standards, Office of Air and Radiation, U.S. EPA. http://www3.epa.gov/ttn/naaqs/standards/ozone/data/20141021welfarerea.pdf

U.S. EPA (Environmental Protection Agency). (2015a). Connectivity of streams and wetlands to downstream waters: A review and synthesis of the scientific evidence. (EPA/600/R-14/475F). Washington, D.C.: National Center for Environmental Assessment, National Exposure Research Laboratory, National Health and Environmental Effects Research Laboratory, Office of Research and Development, U.S. EPA. http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=296414

U.S. EPA (Environmental Protection Agency). (2015b). TSCA work plan chemical risk assessment. N-methylpyrrolidone: paint stripper use. (740-R1-5002). Washington, D.C.: Office of Chemical Safety and Pollution Prevention, U.S. EPA. http://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P100M55I.TXT

URS Greiner Inc.; CH2M Hill. (2011). Coeur d'Alene basin remedial investigation/feasibility study. (URS DCN: 4162500.06200.05.a2). Seattle, WA: Region 10, U.S. EPA. http://yosemite.epa.gov/r10/cleanup.nsf/8065e3cf3d5268538825778300663abc/70761fd9f6ee19c988256cce0006f78f/$FILE/Preface.pdf

van der Ohe, PC; De Zwart, D; Semenzin, E; Apitz, SE; Gottardo, S; Harris, B; Hein, M; Marcomini, A; Posthuma, L; Schafer, RB; Segner, H; Brakck, W. (2014). Monitoring programmes, multiple stress analysis and decision support for river basin management. In J Brils; W Brack; P Negrel; JE Vermaat (Eds.), Risk-Informed Management of European River Basins. Berlin, DL: Springer-Verlag Heidelberg.

Villeneuve, D; Volz, DC; Embry, MR; Ankley, GT; Belanger, SE; Leonard, M; Schirmer, K; Tanguay, R; Truong, L; Wehmas, L. (2014). Investigating alternatives to the fish early-life stage test: a strategy for discovering and annotating adverse outcome pathways for early fish development. Environ Toxicol Chem 33: 158-169. http://www.ncbi.nlm.nih.gov/pubmed/24115264

Wang, NC; Jay Zhao, Q; Wesselkamper, SC; Lambert, JC; Petersen, D; Hess-Wilson, JK. (2012). Application of computational toxicological approaches in human health risk assessment. I. A tiered surrogate approach. Regul Toxicol Pharmacol 63: 10-19. http://www.ncbi.nlm.nih.gov/pubmed/22369873

Weed, DL. (2005). Weight of evidence: a review of concept and methods. Risk Anal 25: 1545-1557. http://www.ncbi.nlm.nih.gov/pubmed/16506981

Woodruff, TJ; Sutton, P. (2014). The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. Environ Health Perspect 122: 1007-1014. http://www.ncbi.nlm.nih.gov/pubmed/24968373

APPENDIX A. GLOSSARY OF WEIGHT-OF-EVIDENCE TERMS

The following terms are defined as used in this document. The U.S. Environmental Protection Agency (EPA) might use these terms differently in other contexts.

Alteration: (1) A characteristic of causal relationships; it specifies that the affected entity is changed by physical, chemical, or other mechanisms leading to the defined effect. (2) A change in an entity that is consistent with interaction with a cause.

Ambiguous: Evidence that has no clear meaning or more than one possible meaning. Evidence may be ambiguous because it is weak (e.g., shows no clear relationships) or unreliable (e.g., reference sites may have important extraneous differences from exposed sites).

Antecedence: A characteristic of causal relationships; it specifies that the causal interaction is itself connected to processes that precede it, potentially leading back to a source.

Assessment, environmental: (1) A process of generating and presenting scientific information to inform an environmental regulatory or management decision. (2) The product of an environmental assessment process.

Benchmark: A criterion, standard, screening value, effect threshold, or other value used to differentiate potential exposure levels that are of concern from levels that are not of concern.

Case: (1) The situation that is the subject of an environmental assessment; for example, the case may be a water body experiencing an algal bloom, a hazardous waste site or a proposed pesticide use. (2) The weighted body of evidence for an assessment hypothesis; for example, "the lack of co-occurrence weakens the case for fine sediments as the cause."

Characteristics: Properties that define a quality of interest and that could be supported by evidence. For example, when weighing evidence for causation, evidence is sought to support six characteristics of causal relationships (Table E-1).

Coherence: The property of a body of evidence that its constituent pieces are logically linked together, thus, forming a reasonable explanation.

Confidence: The credence given a conclusion. For quantitative results, confidence is determined by the scatter of estimates as well as the weight of the evidence.

Considerations: Sets of heterogeneous aspects of causation or other qualitative hypotheses that are used to structure narrative weight of evidence (WoE). They are derived from or analogous to Hill's considerations (Hill, 1965).

Co-occurrence: (1) A characteristic of causal relationships; it specifies that the cause and effect are collocated in space and time at a scale appropriate to the cause and effect. (2) An instance of collocation in space and time.

Correspondence: The similarity of a piece or type of evidence to the entity or conditions to which the evidence will be applied. Evidence is relevant if it corresponds well to all aspects of the case (biological, physical/chemical, and environmental conditions).

Corroboration: Supporting evidence for an assessment proposition from one or more independent studies providing similar results.

Data: Unanalyzed results of measurements, counts, or observations used as a basis for reasoning or calculation.

Discrepant: Evidence that is inconsistent or contrary to established facts or theory.

Endpoint, assessment: An explicit expression of the environmental values to be protected, operationally defined as an ecological entity and its attributes.

Evaluation: The determination of how much influence a piece or category of evidence should be assigned. Evaluation plus the scoring of its results constitutes the weighting of evidence.

Evidence: Information that informs inferences regarding a condition, cause, prediction, or outcome.

Evidence, body of: All the applicable evidence used to make inferences concerning a proposition.

Evidence, category of: A grouping of evidence for consistency and to facilitate weighting. In WoE for environmental assessments, evidence is categorized in terms of conventional *types of evidence*, or in terms of characteristics that the evidence supports.

Evidence, line of: (1) A complex piece of evidence including multiple causal or logical steps that establishes a line of reasoning. (2) A general term used to refer to a piece or type of evidence—this use is discouraged to reduce ambiguous terminology.

Evidence, piece of: The basic unit of evidence; examples include the results of a toxicity test or a stream survey.

Evidence, type of: A category of evidence that is based on a distinct form of study. Conventional types include biological surveys, biomarkers, ambient media toxicity tests, single-substance toxicity tests, population and ecosystem models, and quantitative structure-activity relationships.

Explanation: The process of translating pieces or types of evidence into a characteristic of causation or some other attribute that is relevant to a hypothesis. It is part of evidence integration in the weighing of the body of evidence step.

Hazard: The potentially adverse properties of an agent that, with potential exposure of a receptor, imparts a risk. Hazards are identified by a general causal assessment, such as an Integrated Science Assessment, that typically employs a qualitative WoE.

Hypothesis: A proposition proposed to be a potential explanation of a phenomenon (the fish kill may be caused by high temperatures) or a potential outcome of a phenomenon (building more rooftops and parking areas will interfere with groundwater recharge).

Inference: (1) The act of reasoning from evidence. (2) A result of such reasoning.

Information: Data or other facts used to derive evidence.

Integration: The first step in the weighing of the body of evidence in which the pieces or categories of evidence are combined to characterize the body of evidence.

Interaction: (1) A characteristic of causal relationships; it specifies that a causal agent impinges upon, enters, binds to, or initiates a response in a susceptible entity in a way that can lead to the effect of concern. (2) An instance of interaction of a causal agent on a susceptible entity that can lead to an effect.

Interpretation: The second step in the weighing of the body of evidence in which the body of evidence for each hypothesis is compared to other hypotheses or judged against a standard to determine what conclusion is best supported by the weight of evidence.

Property: A feature of evidence that determines how much weight it should be assigned. In this document, the major properties are relevance, strength, and reliability.

Property, collective: A feature of bodies of evidence that, along with the properties of the constituent pieces of evidence, determines how much weight a body of evidence should be assigned. In this document, the collective properties are number, coherence, diversity, and absence of bias.

Proposition: A condition, cause, prediction, or outcome hypothesized as a possible outcome of an assessment.

Qualitative WoE: Weight of evidence to infer a quality such as causation or impairment.

Quality of interest: A distinctive attribute of an entity, relationship, or system that should be determined to exist or not for a hypothesis in an assessment.

Quantitative WoE: Weight of evidence to estimate a quantity such as a benchmark value or a biodegredation rate.

Reasonable explanation: (1) A statement or account that coherently explains a body of evidence. (2) Informed reasons for apparent inconsistencies in a body of evidence that provide coherence.

Refutation: The logical process of demonstrating the impossibility of a condition, cause, predicted effect, or outcome.

Relevance: A property of a piece or type of evidence that expresses the degree of correspondence between the evidence and the assessment endpoint to which it is applied.

Reliability: A property of evidence determined by the degree to which it has quality or other attributes that inspire confidence**.**

Risk: The likelihood of adverse effects associated with exposure to a stressor.

Scatter: The distribution of measured or estimated values due to uncertainty, variability, or both.

Scoring: The process of formalizing the evaluation of evidence by assigning a numeral, term, or symbol. Evaluation plus scoring constitutes weighting of evidence.

Strength: A property of evidence determined by the degree of differentiation from randomness or from control, background, or reference conditions.

Sufficiency: (1) A characteristic of a causal relationship; it specifies that the causal agent or event must be adequate to induce the effect in susceptible entities. (2) An occurrence of enough of an agent or process to affect a susceptible entity.

Table, scoring: A table created in the weighting step that presents the pieces or categories of evidence and their scores with respect to relevant properties.

Table, weight of evidence: A table created in the weighing step that summarizes the bodies of evidence and their weights for one or more hypotheses.

Time order: (1) A characteristic of a causal relationship; it specifies that the cause precedes the effect. (2) The sequence, in time, of the occurrence of a candidate cause and the effect of concern. It is sometimes called temporality or temporal sequence.

Uncertainty: Lack of knowledge concerning the state of an organism or system or concerning the true value of a quantity. Uncertainty is a property of the investigator and, unlike variability, may be reduced by measurement or observation.

Variability: Heterogeneity over time, space or members of a population. Variability is a property of nature and may not be reduced by measurement or observation.

Weigh: Consider the relevance, strength, and reliability of the body of evidence to assess the likelihood of a hypothesis.

Weight: (noun) The importance to an inference of a piece or category of evidence. (verb) To assign an importance descriptor or score to a piece or category of evidence.

Weight of evidence: (1) A process of making inferences from multiple pieces of evidence, adapted from the legal metaphor of the scales of justice. (2) The relative degree of support for a conclusion provided by evidence. The result of weighing the body of evidence.

APPENDIX B. WEIGHT-OF-EVIDENCE METHODS FOR QUANTITATIVE RESULTS

Assessors are often confronted with multiple estimates of a quantitative value such as the median lethal concentration (LC_{50}) for a fish species, the half-life of a chemical in surface water, or the pH at which species richness is reduced in streams. A single estimate might be derived from multiple estimates, by weighing evidence to identify the best estimate or by using statistical techniques from meta-analysis to combine estimates (Borenstein et al., 2009; Hunter and Schmidt, 2004).

A simple method for combining values is selection of the single best (i.e., weightiest) estimate. This approach is commonly used because often one quantitative estimate is clearly superior to all others, so averaging would diminish accuracy. This approach also might be necessitated by statistical considerations, if the values cannot be combined because they do not represent independent samples from a common population of estimates (Borenstein et al., 2009). When this approach is used, a good practice is to identify in advance the properties of good estimates. Various procedures and criteria can be applied to various study properties when weighting to identify the best value (see Section 5). For example, LC_{50} values might be screened using reliability properties such as use of good laboratory practices and then the most relevant value could be chosen (e.g., the test performed in water most similar to water at the assessment site).

A common method in meta-analysis is averaging of multiple estimates. The geometric mean is often used because many environmental data sets are asymmetrical. Although weighting studies before averaging them is not common in environmental assessments, weighted averages are conventional in meta-analysis. The most common statistical model for weighting is the fixed-effects model, which weights by the inverse of the error variance (Borenstein et al., 2009). The term fixed effects is used because it implies that the estimates are all based on sampling a common variable from a common population with a set effects level (i.e., the only significant source of differences among the estimates being combined is sampling error). It has the advantage of giving more weight to larger studies and of normalizing the variance. If this assumption does not hold, a random effects model or the IVhet model can be employed (Borenstein et al., 2009).

Weighting also might be based on properties of the individual members of a higher-level entity that is being assessed. For example, in an assessment of ozone effects on forest production, estimates of reduction in timber production of individual tree species were weighted by the proportional basal trunk area before they were averaged to estimate loss of forest production (U.S. EPA, 2014b). Because of weighting, the average represented the loss of forest production and not just the averages of production losses for individuals of each species.

Although weighting in meta-analysis is usually based on statistical properties, weights based on qualitative properties also can be used (Turner et al., 2009). This approach is appropriate when studies vary significantly in terms of relevance or reliability relative to sampling variance or variance among the studied systems. For example, chronic toxicity test endpoints, even for the same species, often differ in the endpoint response, so statistical weighting would not capture the most critical differences among chronic test data. A statistical model has been developed that incorporates data quality or any other qualitative scale (Doi and Thalib, 2008). Statistical weighting models depend on variance estimates, which are not available for many environmental values such as conventional toxicity test endpoints. In those cases, weights can be assigned using improvised methods. Weighting for study quality has been common in meta-analysis, but it may introduce bias if not done appropriately (Turner et al., 2009; Greenland and O'Rourke, 2001).

Multiple quantitative estimates also can be used in regression analyses such as derivation of exposure-response relationships from multiple tests. As in averaging, the values are commonly weighted by the inverse of their error variance.

Quantitative results also can be combined into distributions. A common use is derivation of a value in environmental assessments from species sensitivity distributions (SSDs). SSDs are statistical distribution functions that, conventionally, are fit to the results of chemical toxicity tests and are generally interpreted as representing the distribution of sensitivity of species in communities. They are used to derive an exposure level that would be protective of a defined proportion of species or genera in communities or to estimate the proportion of species or genera that a defined exposure level would affect. The derivation of SSDs illustrates the issues involved in quantitative weighting of evidence (Box B-1).

Finally, rather than combining estimates, data sets from multiple studies have been combined to perform a statistical test or analysis. Some investigators combine data sets from multiple small studies with statistically insignificant results to obtain a data set large enough for the differences among treatments to be statistically significant. This practice is *not* recommended because it violates the premise of hypothesis testing that the likelihood of a data set given a null hypothesis is being determined. It is a form of *P*-hacking, which is the selection of data sets, tests, or hypotheses to achieve statistical significance. Alternatively, data sets can be combined to obtain a potentially better estimate of a mean and confidence interval than simply taking the mean of means. Also, multiple sets of exposure-response data might be combined to obtain a more accurate regression model of the relationship. As with averaging estimates, analyzing combined data sets requires attention to the statistical properties of the data.

Choosing the best estimate poses fewer technical issues than combining estimates. Combining estimates or data sets to generate an estimate introduces complications in computation and interpretation due to heterogeneity. In contrast, in a well-designed individual study, what the data and the summary statistics derived from the data represent is known. Different data sets representing the same relationship but generated at different times in different places or at different scales also might be so different due to ecological complexity and variability that they cannot or should not be combined. Further, if the studies to be combined have greatly different sample sizes and variance structures, statistically combining them might give counterintuitive results. At worst, studies that all indicate one relationship (e.g., declining species richness with increasing exposure) might give the opposite result when combined (i.e., increasing species richness with increasing exposure), a phenomenon known as Simpson's paradox (Blyth, 1972). Finally, the set of studies may be biased, particularly by the bias against publishing studies that show no effects. Thus, care should be taken when combining multiple estimates or data sets.

Despite these warnings, the benefits of combining estimates or data sets can be substantial. The combined estimate might be more accurate, particularly when sampling error is important. Inconsistencies among studies can be quantified and their sources can be analyzed. The influence on results of methodological or environmental differences among studies can be quantified. Further, some cases of combining estimates, such as averaging multiple toxicity test results for the same species following a standard protocol with moderate water chemistries, are unlikely to go astray.

Qualitative WoE may be used to derive quantitative results without combining or choosing among estimates. An example is the Rule of Five (Box B-2).

Box B-1. Combining and Weighting Data in Species Sensitivity Distributions

The combining and weighting of data in SSDs has two components. First, to derive the points in the distribution, the results of multiple tests of a species or of closely related species are combined. When deriving the species mean acute value (SMAV) for national ambient water quality criteria, multiple LC_{50}s are combined by taking the geometric mean and assigning that mean value to the species (U.S. EPA, 1985).

$$SMAV = \exp [(\sum \log LC_{50})/n]$$

The genus geometric mean value (i.e., the genus mean acute value or GMAV) is used when multiple congeneric species have been tested.

$$GMAV = \exp [(\sum \log SMAV)/n]$$

Weighting is not part of the currently accepted method for combining test results into either the species or genus means and its implications have not been investigated. However, one might weight the test results (e.g., LC_{50}s) before averaging, based on the variance, number of organisms tested, number of concentrations with partial response or some other property.

$$SMAV = \exp [(\sum(w_t \log LC_{50}))/\sum w_t]$$

Also, one might weight the species mean values before calculating the genus means based on the number of tests of each species or some other property.

$$GMAV = \exp [(\sum w_s \log SMAV)/\sum w_s]$$

(Note: Subscripts *t* and *s* indicate weights [w] for tests and species, respectively.)

Second, weighting might be involved in deriving the SSD from the species or genus mean values. If the SSD is intended to represent aquatic communities, a problem can occur with over- or under-representing particular taxa or functional groups, which might be resolved by weighting the data to achieve proportional representation (Duboudin et al., 2004; Forbes and Calow, 2002; Suter et al., 2002). This situation would require some difficult decisions. Should the weighting be based on the number of species in a higher taxon, a trophic group, or other grouping; on the relative abundance or biomass or on some measure of importance? Invertebrates would be weighted more than vertebrates for species richness, abundance, or biomass and perhaps much less for importance (depending on the measure of importance). Also, the assumption that the weighting should be based on particular categories [e.g., algae, invertebrates, and fish (Duboudin et al., 2004; Forbes and Calow, 2002)] presupposes low variance in sensitivities within those categories relative to between categories. That assumption is questionable given the large taxonomic differences within categories (e.g., cyanobacterial versus eukaryotic algae and arthropods versus mollusks, rotifers, and other invertebrates), which result in large differences in sensitivity. The communities represented by an SSD also would influence the weights. For example, stream invertebrates are predominantly benthic insects, but planktonic crustaceans and rotifers are important in lakes. Clearly, research would be required before introducing a weighting scheme for SSDs.

The derivation of weighted SSDs has been demonstrated by Semenzin et al. (2015); OECD (2013); U.S. EPA (2013). They weighted species and derived the SSD by weighted bootstrapping.

Box B-2. Cleanup Goals by Weight of Evidence Using the Rule of Five

Superfund cleanup goals for ecological risks are concentrations of contaminants that are higher than the no--observed-adverse-effect level (NOAEL) and lower than the lowest-observed-adverse-effect level (LOAEL). Because the interval between these values could be large (i.e., greater than a factor of 10) and because the measures of effect that define the bounds are not consistent, the common practice of interpolating by taking the geometric mean is often inadequate. Instead, a technique called the Rule of Five is used (Greenberg and Charters, 2007; Charters and Greenberg, 2004). The NOAEL-LOAEL interval is divided into six geometrically scaled intervals delimited by seven nodes. The node chosen as the cleanup goal depends on characteristics of the NOAEL and LOAEL, including their relevance to the assessment endpoint and their severity (e.g., Is the LOAEL defined by acute lethality or chronic reproductive effects?). The choice of node also can be influenced by measures of effects other than the LOAEL and NOAEL and by qualitative evidence concerning ecological exposure and effects on the site. The weighing of evidence to choose a node can be, at least in part, rule based. For example, a LOAEL based on mortality would move the choice one node below the middle node. The WoE also uses expert judgment. An example of applying the Rule of Five is in the Baseline Ecological Risk Assessment for the Estuary at the LCP Chemical Site in Brunswick, Georgia.

Another example of using qualitative methods to weigh evidence for quantities is the selection of surrogate values. When there are no data for a property of a chemical, the best estimate of the property may be derived from the best surrogate chemical for which the property is known. This approach, termed "read across," can be based on weighing multiple attributes of the chemicals including structure, physical/chemical properties (e.g., melting point, octanol/water partitioning coefficient) or biological properties [e.g., enzymes induced (ECHA, 2015; Wang et al., 2012)].

APPENDIX C. WEIGHT-OF-EVIDENCE METHODS FOR DERIVING A MODEL

Weight of evidence (WoE) can play three roles in the derivation of mathematical models. First, the selection of a model from a set of alternatives can be based on WoE, where WoE refers to the relative weight provided to a model by the evidence (i.e., the available data). Good (1950) defined the WoE for members of a set of alternative models as their relative likelihoods, given a data set. Models can be compared using the sum of squares or other simple goodness-of-fit statistic, Bayes's factor or Akaike's information criterion (Anderson, 2008; Hilborn and Mangel, 1997). If the models represent alternative causal hypotheses, choice of the model that best explains the data can be said to have chosen the best causal explanation of the modeled effect (Newman and Clements, 2008). This use of the term WoE is distinct from the others in this document. Rather than weighting and weighing multiple pieces or types of evidence, Good's followers weigh the alternative models against each other with respect to a common data set. That approach is included here for completeness because it has appeared in the ecological assessment literature (Linkov et al., 2009; Smith et al., 2002). Rhomberg (2015) described using relative likelihoods for selecting among human toxicological hazards. He admits, however, that it is difficult to calculate likelihoods of models of toxicological hazards, and in practice, he uses qualitative WoE.

Bayesian model averaging extends this concept to provide a more recognizable example of weighing evidence to derive a model (U.S. EPA, 2009b). That is, rather than choosing a single best model, one might combine suitable models after weighting them based on their weights in Good's sense (i.e., their consistency with the data).

A more conventional use of the concept of WoE is the weighing of evidence concerning model assumptions. For example, to determine whether certain phenomena should be included in a model, one might assemble, weight, and weigh evidence concerning their occurrence. Such phenomena might include compensatory processes in a fish population, dietary toxicity in aquatic invertebrates exposed to an aqueous contaminant, nonlinearities in exposure-response relationships, or avoidance behavior at subtoxic exposures. The methods for weighing evidence for these model assumptions would be the same as for other qualitative conclusions. For example, Hertzberg and Teuschler (2002) developed a numerical index to weigh three properties of evidence for alternative models of mixtures toxicity.

Finally, model selection can be based on weighing a mixture of statistical and biological considerations. The benchmark dose guidance recommends choosing a dose-response model based on both biological plausibility and Akaike's information criterion (U.S. EPA, 2012a). A Science Advisory Board panel has recommended broadening the statistical and nonstatistical considerations to be weighed during model selection and using a more formal and transparent weighing process (SAB, 2015). For example, for dose-response modeling, they recommended prioritizing regression models that directly use exposure data for individuals.

Because any causal hypothesis, in theory, can be represented by a mathematical model, uses of WoE for model selection could be equivalent to the approaches for deriving qualitative conclusions presented in Sections 3−7.

APPENDIX D. WEIGHT-OF-EVIDENCE APPROACHES FOR QUALITATIVE CONCLUSIONS

This appendix summarizes the diversity of weight-of-evidence (WoE) approaches that have been used in assessments to answer questions concerning environmental qualities. Each approach has been useful in some cases, and they were carefully considered during the development of the approach presented in this document.

D.1. Narrative Weight of Evidence

Traditional narrative literature reviews and the assessments that adopt their approach describe a body of evidence and reach a conclusion by methods and criteria that are not explicit. The expertise of the reviewer is assumed sufficient to ensure a correct conclusion. The WoE is the conclusion that the author reaches through reading the literature and presents to the reader by logically structuring the review's narrative. This is a common approach for weighing evidence in environmental assessments. Narrative WoE is highly flexible, but the method by which results are obtained is not as transparent or reproducible as for other methods.

D.2. Consideration-Guided Narrative Weight of Evidence

WoE narratives can be given greater structure and logical consistency by organizing the narrative in terms of a set of considerations. For example, the U.S. Environmental Protection Agency's (EPA's) Integrated Risk Information System and Integrated Science Assessment documents use modifications of Hill's considerations when weighing evidence of general causation. Also, the WoE method for Tier 1 screening of potentially endocrine-disrupting chemicals provides five considerations (U.S. EPA, 2012b), and the chemical risk assessments conducted pursuant to the Toxic Substances Control Act use seven considerations (U.S. EPA, 2015b). However, these WoE methods depend on narratives and have not explicitly weighted or tabulated the evidence for the considerations. Also, the considerations are mixtures of various types of issues to consider, not consistent sets of criteria, properties, types of evidence, or characteristics. This method is an advance over unstructured narratives and is generally accepted.

D.3. Evidence Summary and Scoring Tables

Methodological rigor and transparency can be increased by creating summary tables for the evidence concerning an inference and assigning scores. For example, a set of summary and scoring tables was developed at ORNL for contaminated site assessments (Suter et al., 2000). The authors included tables for summarizing the methods and results for each type of evidence: biological surveys, bioindicators, ambient toxicity tests, tissue analyses, and single-chemical toxicity. Another table combined the results of the types of evidence (a score and associated explanation) and presented a score for the body of evidence for an endpoint.

D.4. Criteria-Guided Scoring

Criteria for weighting evidence have been combined with evidence scoring tables in the *Stressor Identification Guidance Document* and Causal Analysis/Diagnosis Decision Information System (CADDIS) to determine the causes of specific environmental impairments (http://www3.epa.gov/caddis/). Several examples of the application of this approach can be found in the CADDIS case studies. Pieces or types of evidence were scored based on Susser's method (U.S. EPA, 2000; Susser, 1986), but scoring types of evidence has been supplemented by scoring characteristics of causation to explain the relationship of the evidence to the hypotheses (Norton et al., 2014; Cormier et al., 2010). This approach

also has been applied to a risk assessment. The Bristol Bay watershed assessment used tables of evidence and scores to clarify the WoE for risks to salmon from spills of tailings, product concentrate and diesel fuel (U.S. EPA, 2014a). Pieces of evidence were described and scored for logical implication (relevance), strength, and quality (reliability), and summary scores were derived for the bodies of evidence. A version of criteria-guided scoring is presented in this document that generalizes the approach beyond causal assessment.

D.5. Standardized Scoring and Weighting

Consistency can be increased by specifying the evidence and criteria to be considered and the scores and weights to be assigned to the possible outcomes for each type of evidence. An example is provided by the Massachusetts weight-of-evidence method for assessing ecological risks at contaminated sites (Menzie et al., 1996). The method numerically scores evidence (termed measurement endpoints) on a scale of 1 to 5 for 10 properties, weights the scores by scaling values based on the importance of the properties, normalizes the scaled scores, and sums the result to give a "measurement endpoint weight." The body of evidence is then qualitatively weighed based on concurrence of the numerical weights. The standardization of scores and weights requires consistency of the cases to which the method is applied and is facilitated by consensus of all parties.

D.6. Indices

Standardized numerical scoring and weighting of evidence has been extended to the calculation of rather complex risk indices (Benedetti et al., 2012; Dagnino et al., 2008; Semenzin et al., 2008; Bombardier and Blaise, 2000). These indices are not quantitative risk estimates, but they often contain component calculations that are part of quantitative risk assessments such as sums of toxic units, so that the index values provide estimates "of relative hazard for the sites being investigated" (Bombardier and Bermingham, 1999). Because the indices often combine evidence across pieces, types, and endpoints, many of them are computationally complex and require support systems for data entry and computations (Semenzin et al., 2009).

Indices are also used to combine multiple metrics from biological survey data into a value that can be said to represent biological integrity (Karr et al., 1986) or deviation from reference condition (Hawkins, 2006). These indices are arithmetic combinations of metrics believed to discriminate impaired and unimpaired waters. Their derivation is not normally considered to constitute a weight-of-evidence process, but they do combine multiple pieces of evidence to determine a quality—the impairment of a biotic community.

Indices are specialized and their results are difficult to relate to the input data. For example, an index of biotic integrity serves its purpose of identifying impaired waters by numerically aggregating numerous incommensurable biological metrics. To assess the cause of the impairment or risks from remedial actions, however, the individual metrics should be extracted, weighted, and weighed appropriately, as described in Sections 4–6.

D.7. Rule Based

If the potential evidence results are few and all the results are reliable, specifying an interpretation of each potential body of evidence can be possible. The decision matrix for the sediment quality triad is a commonly used example of this approach to weighing evidence [Table D-1 (Chapman, 1990)]. The triad is used to determine whether toxic contaminants are impairing a sediment community. Its elements are chemical analyses of sediments that are compared to toxicological benchmark values, toxicity tests of the sediment and surveys of the sediment biotic community. If the scores for the triad elements are, for example, −, +, +, "unmeasured toxic chemicals are causing degradation," but if it is −, −, +, "alteration is

not due to toxic chemicals." The logic of Chapman's decision matrix is correct if the evidence is reliable, but the evidence might not be sufficiently reliable to support it for many reasons (Table D-2). As a result, weighting and interpreting the evidence often is required. Thus, the standard decision matrix is an idealization that serves primarily to clarify the potential relationships among types of evidence. More recently, ambiguities in results have been recognized and new decision matrices have been developed for sediment assessments that specify a decision for only a few of the possible outcomes; for the other outcomes that include conflicting results, they direct assessors to perform additional analyses (COA Sediment Task Group, 2008; Grapentine et al., 2002). Those additional analyses could use any of the other qualitative WoE methods.

Table D-1. Inference based on the sediment quality triad (Chapman, 1990)

Situation	Chemicals Present	Toxicity	Community Alteration	Possible Conclusions
1	+	+	+	Strong evidence for pollution-induced degradation
2	−	−	−	Strong evidence for no pollution-induced degradation
3	+	−	−	Contaminants are not bioavailable, or are present at nontoxic levels.
4	−	+	−	Unmeasured chemicals or conditions exist with the potential to cause degradation.
5	−	−	+	Alteration is not due to toxic chemicals.
6	+	+	−	Toxic chemicals are stressing the system but are not sufficient to significantly modify the community.
7	−	+	+	Unmeasured toxic chemicals are causing degradation.
8	+	−	+	Chemicals are not bioavailable or alteration is not due to toxic chemicals.

Responses are shown as either positive (+) or negative (−) indicating whether measurable and potentially significant differences from control/reference conditions are determined.

Table D-2. Some ways in which the triad decision table might fail (not including errors or poor technique)

Type of Evidence	False Positive (Exaggerated Effect)	False Negative (Minimized Effect)
Single-chemical analyses/ toxicity tests	Higher bioavailability in the laboratory Site background is high Overly sensitive test organisms (sensitive test species or life stages not present at the site)	Short duration Insensitive test species or life stages Critical response not included Multiple chemicals interact Incomplete routes of exposure
Whole-medium samples/ toxicity tests	Bioavailability increased by handling Sampling below bioactive zone Overly sensitive test organisms Inappropriate reference sites	Chemical lost or altered by handling Insensitive test organisms Inappropriate reference sites Insensitive test organisms Ambient conditions (e.g., UV light) are important to effects Critical response not included Contaminated locations or times missed
Sampling locations/ biological surveys	Confounding Inappropriate reference sites	Organisms too rare Response too infrequent Sensitive organisms not sampled Critical response not included Large sample scale Contaminated locations missed Inappropriate reference sites

Another rule for weighing evidence is independent applicability. The Office of Water classifies a water body as impaired if an adequate piece of evidence demonstrates an impairment, even if other evidence did not detect the impairment. This policy takes into consideration the possibility of false negatives and that different types of evidence detect different types of impairment (Box D-1).

A two-stage rule was developed by the European Center for Ecotoxicology and Toxicology of Chemicals for weighing human and animal data in chemical risk assessments (ECETOC, 2009). The category of evidence with the highest quality for a chemical is used, but if the quality of animal and human evidence is similar, an effect is assumed to be caused by the chemical if either category of evidence shows the effect. This logic could be applied to more than two categories of evidence if equivalent quality scales could be developed for all categories.

D.8. Quantitative Alternatives

Some authors have suggested that quantitative methods are inherently superior for making qualitative inferences from multiple pieces of evidence (Linkov et al., 2009). In particular, multi-criteria decision analysis (MCDA) and Bayesian belief networks (BBNs) have been suggested to be alternatives to conventional WoE (Linkov et al., 2015; Fenton and Neil, 2013).

Box D-1. Independent Applicability and Weight of Evidence

Independent applicability of evidence of impairment has been a policy of the EPA Office of Water. That is, if a water body fails to meet water quality criteria, is toxic, or is biologically impaired, it is an impaired water body. Independent applicability can result from two possible logics.

1. All types of evidence are fallible, so any evidence can generate false negative results. Rather than weigh evidence of impairment against evidence that failed to detect impairment, the evidence of impairment is accepted. Therefore, if any type of reliable evidence detects impairment, the water body is considered impaired. This logic is defensible if the consequences of not detecting an impaired ecosystem are more serious than the consequences of failing to detect. This justification is policy based. In addition, the logic can be defended by recognizing that poor methods or poor implementation are inherently more likely to fail to detect an effect than to detect a false effect. This justification is based on the fact that poor practices introduce extraneous variance are not spatially comprehensive and otherwise tend to obscure effects. Further, it is unlikely that a type of evidence will test or measure the most sensitive conditions, taxa, life stages, responses and interactions between species and stressors. Therefore, even good studies are likely to miss real effects. This does not imply that false positives never occur (see Table D-1), but rather, poor evidence is more likely to result in false negatives and unprotective decisions.

2. Each type of evidence addresses a distinct aspect of impairment. If any one aspect is found in a water body, the water body is impaired in that way. This assumption involves no synthesis. Each type of evidence is independently evaluated with respect to its aspect of impairment. This is explicitly the case with the European Union's Water Framework Directive that distinguishes good chemical status from good ecological and good hydrological statuses (van der Ohe et al., 2014; EC, 2013). European waters should achieve all three. Equivalently, the policy of independent applicability can be interpreted as a requirement to protect three types of integrity: chemical, toxicological, and ecological. This approach is supported by the inherent differences in the types of evidence. For example, a biocriterion based on the index of biotic integrity discriminates sites on a human disturbance gradient (Blocksom and Johnson, 2009). The index responds to the multiple stressors that occur commonly on such gradients and is most responsive to those stressors that are common components of human disturbance. Many of those stressors, such as flashy flow, have no numeric criteria. On the other hand, water quality criteria are based on responses to particular chemicals or other stressors and not necessarily to common human disturbances of aquatic systems. Water quality criteria and biocriteria are criteria for different impairments.

Note that this discussion applies to integrating evidence across types. When applying independent applicability, one might weigh evidence within types. For example, multiple toxicity tests of an ambient water could be combined to determine whether the water has unacceptable toxicity. Also, once a water body is declared impaired, WoE approaches can be applied when assessing the risks associated with alternative remedial or regulatory actions.

Although WoE integrates evidence concerning some environmental quality of interest, MCDA is intended to go farther and identify the best decision, given the problem framing, decision criteria, and preferences [e.g., Which is the best option for dredge spoil disposal given costs, risks, and stakeholder preferences? (Stahl et al., 2002)]. MCDA conventionally requires that probabilities be estimated for the outcomes of a decision and that a decision maker assign utilities to attributes of the outcomes (i.e., the decision criteria) to allow calculation of the expected utility of a decision (Linkov and Moberg, 2012). Weighting is involved in conventional MCDA when combining the multiple decision criteria (cost, aquatic toxicity, public acceptance, etc.). The computational methods of MCDA, however, have been adapted for a variety of uses beyond decision making, including combining numerically weighted evidence (Linkov and Moberg, 2012). One example is weighing the evidence for adverse outcome pathways [Box 3-3, (Collier et al., 2016)]. Linkov et al. (2015) described MCDA as a suitable proxy for Bayesian analysis.

BBNs are graphical models of the uncertainties and dependencies used to calculate the probability of an outcome in a causal network (Fenton and Neil, 2013). BBNs adapted for decision analysis are termed influence diagrams (Carriger and Barron, 2011). BBNs are computationally complex and not familiar to most environmental assessors and stakeholders, but available software makes their application relatively easy. When BBNs are involved in environmental inferences, they may not have data-derived probabilities; instead expert judgment is commonly used. These networks are WoE in the sense that the probability of a variable's state can be considered as its weight and branches of the network can considered chains of evidence (Small, 2008). For example, the probabilities of fishery closure could be one weighted piece of evidence and probabilities of levels of ecological impacts might be another; these can be combined by calculating the conditional probabilities to derive levels of offshore ecosystem services impacts (Carriger and Barron, 2011). Carriger and Barron (2016) have adapted BBNs to explicitly weigh multiple types of ecological evidence. They treated the three types of evidence in the sediment quality triad as branches of a BBN, calculated probabilities that each type was indicative of effects, weighted each in terms of its importance, and then calculated an overall probability of impairment. Because this method does not use a causal network, it is no longer a BBN in the conventional sense. However, it does provide a means of computationally combining evidence rather than using qualitative logic.

Although these methods have not been used for conventional qualitative WoE tasks such as identifying causes or hazards, nothing in this document precludes using them where appropriate. An example of a relevant potential regulatory application is the use of BBNs to optimize choices of toxicity tests in European chemical regulation in place of judgment-based WoE (Jaworska et al., 2010). This case is ideal for these quantitative methods because it is a relatively simple and consistent problem with a clear decision structure. Objectively estimating constituent probabilities of, for example, a positive rat carcinogenicity test given a positive Ames assay is possible. One simply needs a sufficient database of test results.

D.9. Summarize Only

Some assessors have argued that the weighing of evidence should not determine what conclusion is best supported by the evidence (Gray, 1994). Instead, they argue that assessors should assemble, clearly summarize, and simply present the evidence to the decision maker with enough background information for them to make a decision. Although this approach takes subjective judgments concerning the integration of evidence away from assessment scientists, it confers responsibility on decision makers to interpret scientific information, which they might not be trained to do, and requires time for inferential work that they might not have.

D.10. Conclusions

This summary of WoE approaches is intended to show the range of techniques that might be used for qualitative WoE. Methods proposed or used, however, do not necessarily fit into these categories. For example, McDonald et al. (2007) translated qualitative weights to numbers that are combined algebraically and then translated into a final qualitative conclusion. This method is typical of the WoE literature, which is replete with methods developed for the individual application. The inclination to develop a new WoE method for each application might be diminished by the availability in this document of a generic WoE approach that is flexible but provides some consistency across a variety of applications.

APPENDIX E. CHARACTERISTICS OF INFERRED QUALITIES

Qualitative weight of evidence (WoE) addresses qualities, each of which is characterized by certain characteristics. For example, if the quality of interest is causation, the hypothesized relationship should have the characteristics of causation such as, the causal agent and the affected entity interact. Association of evidence with characteristics of causation, or characteristics of any other quality being addressed, can serve two purposes. First, characteristics explain what the evidence does for the hypothesis (Section 6.2). For example, aqueous concentrations of a chemical in streams can demonstrate co-occurrence but not interaction. Second, categorizing the evidence in terms of the characteristics can help to perform and justify the inference. For example, a WoE table organized in terms of causal characteristics may show that there is abundant evidence for co-occurrence but none for interaction, which would raise questions about bioavailability. Only characteristics of causation are established, but considering appropriate characteristics of other inferred qualities such as impairment or protectiveness also might be useful. Potential characteristics for five qualities are presented here.

E.1. Characteristics of Causation

In environmental assessments, WoE has been used primarily to assess causality. In fact, WoE has been equated with determining the degree of credence due a causal hypothesis (Rhomberg et al., 2013; Krimsky, 2005). WoE is particularly associated with causation because causality is a fundamental concept, and no one piece or type of evidence can prove causation (Norton et al., 2014). Rather, the relationships of evidence to characteristics of causal relationships should be considered. Six characteristics of causation, developed for ecological causal assessments, have proven useful for that purpose (Table E-1). By analogy to Hill's considerations, they have been called Cormier's causal characteristics (CCC).

Few bodies of evidence for causal assessments include evidence for all of these characteristics. Even in very well-studied cases, evidence of conditions prior to induction of the effect is seldom available to consider time order (Table 7-1). Even an incomplete WoE table based on characteristics, however, provides more insight into what is known about causation in a case than a table based only on types of evidence.

Although these characteristics were developed for causation in specific cases, they also can be relevant to general causation (i.e., Does agent *x* cause effect *y*?), if actual cases are available. The characteristics can be used to demonstrate that, at least in those actual cases, the agent could cause the effect. They can be particularly useful when the general question is framed in terms of real-world conditions. An example is the use of the characteristics to determine that, where aqueous conductivity is elevated due to major mineral ions in West Virginia and Kentucky, the ion mixture causes the extirpation of invertebrate genera (Cormier et al., 2013; Cormier and Suter, 2013; U.S. EPA, 2011b).

Although the characteristics apply to, and can be observed in actual cases, they are not observable in the same way when the general causal question is hypothetical. That is, hypothetically, can *x* cause *y* rather than, in fact, did *x* cause *y* or is *x* causing *y*? In particular, when the only evidence of effects is from toxicity tests, the co-occurrence and its antecedents are simply the design and setup of the test, and the time order is inevitable. Sufficiency in hypothetical causal assessments is not of interest because no particular effect level is identified, so no particular exposure is sufficient. Instead, we use the test data to establish an exposure-response relationship that can be applied generally. The real causal concerns are with the relevance of the co-occurrence, interaction, and alteration in the test to those in the field. Are the conditions, durations, species, and life stages in the tests relevant to the co-occurrence of interest in the

field? Are the interactions of the stressor and test organisms and the resulting alterations plausible in the field? Are others likely to be more sensitive or important?

Table E-1. Characteristics of causal relationships[a]

Causal Characteristics	Description	What Evidence of a Characteristic Shows
Time order	The cause precedes the effect.	Change in the entity after exposure to the cause and not before
Antecedence	The causal relationship is a result of a larger web of antecedent cause-and-effect relationships. Evidence includes sources and routes of transport.	Earlier events that led to the particular causal event. It demonstrates that a potential causal event is part of a network of prior causal events, which increases confidence that the potential causal event actually occurred (e.g., that it was not a result of a measurement error or hoax).
Co-occurrence	The cause co-occurs with the susceptible entity in space and time. Evidence includes contaminant concentrations in ambient media and habitat alterations.	The presence of both a causal agent and the affected entity in circumstances that create the potential for exposure
Interaction	The cause interacts with the entity in a way that can lead to the effect. Evidence includes injuries, body burdens, and biomarkers.	Signs that the entity has been exposed to the agent in such a way that changes can be initiated
Sufficiency	The intensity, frequency and duration of the cause are adequate, and the entity is sufficiently susceptible to produce the type and magnitude of the effect.	Enough of the cause and a sufficiently susceptible entity that can result in the level of the observed effect
Alteration	The entity is changed by physical, chemical, or other mechanisms leading to the defined effect. Evidence includes symptoms and related attributes of the affected entities.	Changes in the entity attributable to or at least appropriate to the cause, which can indicate that the mechanism is acting

[a] Modified from Norton et al. (2014); Cormier et al. (2010).

The characteristics of causation can be arranged in a logical sequence: *antecedent* causes lead to *co-occurrence* of the proximate cause with susceptible or affected organisms, which leads to *interaction* with the affected organisms and, if the interaction is *sufficient*, leads to *alteration* of the organisms. The characteristics also serve to prompt assessors to make full use of the evidence with respect to all characteristics of causation. For example, it is obvious that laboratory test data provide evidence concerning sufficiency, but the data also might provide evidence of alterations characteristic of organisms affected by the tested agent that might be observed in the field.

E.2. Characteristics of Protection

Criteria, standards, screening levels and other benchmarks are intended to define a limit on exposure that would be adequately protective. Protection, like causation, is a difficult concept with multiple characteristics (Table E-2). Benchmarks are usually derived using one piece or one type of evidence. WoE might be used, however, to assess the confidence provided by diverse evidence that a benchmark or a procedure for deriving benchmarks is appropriately protective. Although the U.S. Environmental Protection Agency has not explicitly used this potential application of WoE, characteristics like those Table E-2 might be useful.

Table E-2. Potential characteristics of a protective benchmark

Characteristics of Protective Benchmarks	Description
Causal relationship	The exposure-response relationship used in the derivation is likely to be causal.
Predictive model of the relationship	The model of the relationship predicts the effects of specified exposures.
Represent sensitive endpoints	Known sensitive endpoint species or other entities and responses can be represented by the relationship.
Exposure metric is relevant	The exposure metric represents the relevant exposure of the sensitive entities.
Set at an appropriately protective level	Independent data demonstrate, with sufficient confidence, a lack of effects of concern on the aquatic community and constituent taxa.
Discriminatory power	The benchmark discriminates between impaired and unimpaired communities with sufficient confidence.

E.3. Characteristics of Contaminants of Concern

Risk assessments of contaminated sites typically develop lists of contaminants of concern based on chemical analyses of site media. The contaminants of concern are then the subjects of the risk assessment. Concern may be determined by comparison to ecological benchmarks, but more often multiple characteristics of a contaminant are weighed to determine which make the list. The characteristics in Table E-3 are based on a text and a recent Superfund ecological risk assessment (Black & Veatch Special Projects Corps, 2011; Suter et al., 2000). The list may not be exhaustive.

Table E-3. Characteristics of a contaminant of concern at a contaminated site

Characteristics of Contaminants of Concern	Description
Associated with the waste or source	Chemicals that are not associated with the waste or effluent of interest could be from an extraneous local source, regional contamination, or natural background.
Frequently detected	Chemicals that are infrequently detected at the site are of lesser concern and might be artifacts.
Greater than local background	Chemicals at background levels are not of concern, unless the anthropogenic form is more bioavailable or toxic.
At potentially toxic levels	Concentrations above relevant ecotoxicological benchmarks are of concern.
Detection limits are inadequate	Chemicals that could not be detected at toxic levels are of concern, particularly if they are associated with the waste or source.
Belongs to the same class as an identified contaminant of concern	A chemical that is not of concern alone might belong to a class, such as polycyclic aromatic hydrocarbons, that could be assessed jointly.

E.4. Characteristics of Biological Impairment

Impairment of biotic populations or communities is determined under the Clean Water Act's Section 303(d) and in various other contexts. Although a population or community need be impaired in only one way to be categorized as impaired (Box D-1), multiple characteristics of impairment could increase confidence in the designation and increase support for a regulatory or remedial action. A few of

the possible characteristics of impairment include low taxonomic richness, proportion of dominant taxa, toxicity, contamination, poor physical habitat, and poor aesthetics, among others. These characteristics are familiar to ecological assessors and require no descriptions here.

E.5. Characteristics of Remediation

Assessments of the success of remedial actions are common and important examples of outcome assessments. Remediation is judged in terms of performance and effectiveness of the intervention (Table E-4). Obtaining sufficient evidence to determine whether ecological goals are attained because of the remedial action and not just natural recovery is valuable.

Table E-4. Potential characteristics of remediation

Characteristic of Remediation	Description
Performance of the Intervention	
Functions	The remedial technology functions as intended.
Reduces the agent	The technology substantially reduces the level and extent of the contaminant or other agent that causes the risk.
Achieves the physical/chemical goal	The levels of the causal agent are reduced sufficiently to achieve the goal.
Effectiveness of the Intervention	
Endpoint less exposed	The exposure of the environmental endpoint entity to the agent and their interaction are reduced as predicted.
Endpoint improved	The abundance, diversity, function, or other attributes of the endpoint entity improve within the expected time following the remediation.
Endpoint goal achieved	Reference conditions or other ecological endpoint goal are achieved in response to remediation.

E.6. Summary

Qualitative WoE is performed to support inferences concerning some qualitative property such as causality, protection, concern, or impairment. Causality has been the principal quality inferred by WoE in environmental assessments and is the only quality for which characteristics have been available. Tentative lists of characteristics are presented here as examples to demonstrate the potential use of characteristics as an organizing principle for WoE. Other qualitative properties such as recovery also might be addressed by weighing the evidence for defined characteristics. As in causal inference, weighing evidence for characteristics of these other properties could provide a better explanatory structure to WoE than simply weighing pieces or types of evidence.