Weight of Evidence
Cheat Sheet

1. **Data Quality** (QAPP) - Whether data is considered adequate or inadequate? Was there a QAPP associated with the data gathering process?

2. **Linkage** – Extent to which measurement endpoint links to possible BU impacts.

   - High linkage example: direct measurement of a constituent concentration exceeding established water quality objective designed to protect a specific BU. Measurements that can be used to be compared with CTR, NTR, and TIE values.
   - Medium linkage – measurements that can be used to indirectly link to BU impact such as MTRLs for human fish consumption.
   - Low Linkage – The link of the measurement is weak such as using biomakers to assess biological impact to biological communities.

3. **Utility** of measurement- Extent to which a well-accepted standards, criteria, guidelines or other objective measurements determining if a narrative water quality objective is met and a BU is attained.

   - High utility measurements – NAS values, FDA action levels, U.S.EPA screening values, Maximum contaminant levels (MCLs), fish advisories, BPTCP approaches, beach closures, and postings, published temperature thresholds, published sedimentation thresholds, Federal agency and other state sediment quality guidelines, DHS bacteria standards, DG guidelines. Any adopted numerical WQO, CTR or NTR values for a narrative objective.
   - Medium utility - Accepted measurement to determine if a narrative is being met but the applicability is limited and the scientific basis is weak or insensitive. These include Sediment apparent Effects thresholds from other states. MTRLs.
   - Low utility – Measure is limited and has limited applicability and certainty.

4. **Water Body Specific Information** –factors associated with the measurement at Water body segment of interest which increases the confidence of the monitoring data. These include age of data (10 years or less), environmental data measured at the site, Species or indicator present, environmental conditions at site: seasonality, storms, land use practice, etc.

   - High – 4 factors reflect the site.
   - Medium - 2 factors reflect the site.
   - Low – one factor reflect the site.

5. **Spatial representativness** – Relates to the degree of compatibility or overlap between the study area, locations of measurement or samples, locations of stressors, and locations of ecological receptors and their potential exposure.

6. **Temporal representivness** – relates to the temporal compatibility or overlap between the measurement endpoint (when data was collected).

7. **Quantitativeness and sensitivity of the measurement** - relates to whether the data is or is not numerical. The ability to detect an acute or chronic response in the measurement expressed as the number of samples or occurrences to determine if a WQO has been exceeded.

8. **Use of standard methods** – refers to the extent to which the study followed standard protocols.

§ 1362. Definitions <502>

(6) The term **"pollutant"** means dredged spoil, solid waste, incinerator residue, sewage, garbage, sewage sludge, munitions, chemical wastes, biological materials, radioactive materials, heat, wrecked or discarded equipment, rock, sand, cellar dirt and industrial, municipal, and agricultural waste discharged into water. This term does not mean (A) "sewage from vessels" within the meaning of section 312 of this Act [33 USCS § 1322]; or (B) water, gas, or other material which is injected into a well to facilitate production of oil or gas, or water derived in association with oil or gas production and disposed of in a well, if the well used either to facilitate production or for disposal purposes is approved by authority of the State in which the well is located, and if such State determines that such injection or disposal will not result in the degradation of ground or surface water resources.

(12) The term **"discharge of a pollutant"** and the term "discharge of pollutants" each means (A) any addition of any pollutant to navigable waters from any point source, (B) any addition of any pollutant to the waters of the contiguous zone or the ocean from any point source other than a vessel or other floating craft.

(13) The term **"toxic pollutant"** means those pollutants, or combinations of pollutants, including disease-causing agents, which after discharge and upon exposure, ingestion, inhalation or assimilation into any organism, either directly from the environment or indirectly by ingestion through food chains, will, on the basis of information available to the Administrator, cause death, disease, behavioral abnormalities, cancer, genetic mutations, physiological malfunctions (including malfunctions in reproduction) or physical deformations, in such organisms or their offspring.

(14) The term **"point source"** means any discernible, confined and discrete conveyance, including but not limited to any pipe, ditch, channel, tunnel, conduit, well, discrete fissure, container, rolling stock, concentrated animal feeding operation, or vessel or other floating craft, from which pollutants are or may be discharged. This term does not include agricultural stormwater discharges and return flows from irrigated agriculture.

(16) The term **"discharge"** when used without qualification includes a discharge of a pollutant, and a discharge of pollutants.

(19) The term **"pollution"** means the man-made or man-induced alteration of the chemical, physical, biological, and radiological integrity of water.

5%

USE THIS TABLE WITH THE STANDARD RECOMMENDATION CHEAT SHEET FOR P=.05

Samp size	Confidence to List				Confidence to De-list			
	0.9	0.8	0.7	0.65	0.1	0.2	0.3	0.35
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0
6	1	1	0	0	0	0	0	0
7	1	1	1	0	0	0	0	0
8	1	1	1	0	0	0	0	0
9	1	1	1	1	0	0	0	0
10	1	1	1	1	0	0	0	0
11	2	1	1	1	0	0	0	0
12	2	1	1	1	0	0	0	0
13	2	1	1	1	0	0	0	0
14	2	1	1	1	0	0	0	0
15	2	1	1	1	0	0	0	0
16	2	1	1	1	0	0	0	0
17	2	2	1	1	0	0	0	0
18	2	2	1	1	0	0	0	0
19	2	2	1	1	0	0	0	0
20	2	2	1	1	0	0	0	0
21	2	2	1	1	0	0	0	1
22	2	2	2	1	0	0	0	1
23	3	2	2	1	0	0	0	1
24	3	2	2	1	0	0	1	1
25	3	2	2	2	0	0	1	1
26	3	2	2	2	0	0	1	1
27	3	2	2	2	0	0	1	1
28	3	2	2	2	0	0	1	1
29	3	2	2	2	0	0	1	1
30	3	2	2	2	0	0	1	1
31	3	3	2	2	0	0	1	1
32	3	3	2	2	0	1	1	1
33	3	3	2	2	0	1	1	1
34	3	3	2	2	0	1	1	1
35	3	3	2	2	0	1	1	1
36	4	3	2	2	0	1	1	1
37	4	3	2	2	0	1	1	1
38	4	3	2	2	0	1	1	1
39	4	3	3	2	0	1	1	1
40	4	3	3	2	0	1	1	1
41	4	3	3	2	0	1	1	1
42	4	3	3	3	0	1	1	1
43	4	3	3	3	0	1	1	1
44	4	3	3	3	0	1	1	2
45	4	3	3	3	1	1	1	2
46	4	3	3	3	1	1	1	2
47	4	4	3	3	1	1	1	2
48	4	4	3	3	1	1	1	2
49	4	4	3	3	1	1	2	2

50	5	4	3	3	1	1	2	2
51	5	4	3	3	1	1	2	2
52	5	4	3	3	1	1	2	2
53	5	4	3	3	1	1	2	2
54	5	4	3	3	1	1	2	2
55	5	4	3	3	1	1	2	2
56	5	4	4	3	1	1	2	2
57	5	4	4	3	1	1	2	2
58	5	4	4	3	1	1	2	2
59	5	4	4	3	1	2	2	2
60	5	4	4	4	1	2	2	2
61	5	4	4	4	1	2	2	2
62	5	4	4	4	1	2	2	2
63	5	5	4	4	1	2	2	2
64	6	5	4	4	1	2	2	2
65	6	5	4	4	1	2	2	2
66	6	5	4	4	1	2	2	2
67	6	5	4	4	1	2	2	3
68	6	5	4	4	1	2	2	3
69	6	5	4	4	1	2	2	3
70	6	5	4	4	1	2	2	3
71	6	5	4	4	1	2	2	3
72	6	5	4	4	1	2	3	3
73	6	5	5	4	1	2	3	3
74	6	5	5	4	1	2	3	3
75	6	5	5	4	1	2	3	3
76	6	5	5	4	1	2	3	3
77	6	5	5	4	2	2	3	3
78	6	5	5	5	2	2	3	3
79	7	6	5	5	2	2	3	3
80	7	6	5	5	2	2	3	3
81	7	6	5	5	2	2	3	3
82	7	6	5	5	2	2	3	3
83	7	6	5	5	2	2	3	3
84	7	6	5	5	2	2	3	3
85	7	6	5	5	2	3	3	3
86	7	6	5	5	2	3	3	3
87	7	6	5	5	2	3	3	3
88	7	6	5	5	2	3	3	3
89	7	6	5	5	2	3	3	4
90	7	6	5	5	2	3	3	4
91	7	6	6	5	2	3	3	4
92	7	6	6	5	2	3	3	4
93	7	6	6	5	2	3	3	4
94	7	6	6	5	2	3	3	4
95	8	6	6	5	2	3	4	4
96	8	7	6	5	2	3	4	4
97	8	7	6	6	2	3	4	4
98	8	7	6	6	2	3	4	4
99	8	7	6	6	2	3	4	4
100	8	7	6	6	2	3	4	4

Assessment
Methodology

DRAFT REPORT
A WEIGHT-OF-EVIDENCE APPROACH FOR EVALUATING ECOLOGICAL RISKS

Prepared by

Massachusetts
Weight-of-Evidence Workgroup

November 2, 1995

ACKNOWLEDGEMENTS

The Weight-of-Evidence Workgroup

Nancy Bettinger, Massachusetts Department of Environmental Protection (DEP), Boston, Massachusetts

Jerry Cura, Menzie-Cura & Associates, Chelmsford, Massachusetts

Ken Finkelstein, NOAA Coastal Resource Coordinator, Boston, Massachusetts

Jack Gentile, University of Miami, Miami, Florida (formerly with U.S. Environmental Protection Agency Risk Assessment Forum)

Miranda Hope Henning, McLaren/Hart - ChemRisk, Portland, Maine

Jamie Maughn, CH2M HILL, Boston, Massachusetts

Charlie Menzie (Chairperson), Menzie-Cura & Associates, Chelmsford, Massachusetts

Dave Mitchell, ENSR, Acton, Massachusetts

Stephen Petron, CH2M HILL, Boston, Massachusetts (formerly with Metcalf & Eddy, Wakefield MA)

Bonnie Potocki, CDM, Federal Inc., Boston, Massachusetts

Sue Svirsky, U.S. Environmental Protection Agency, Boston, Massachusetts

Patti Tyler, U.S. Environmental Protection Agency, Lexington, Massachusetts

Other Support

Meetings were held either at Metcalf & Eddy or at Menzie-Cura & Associates. ChemRisk and Menzie-Cura & Associates provided logistical support.

Helpful comments were received from Rick Suggat (Normandeau Associates) and several staff members at Massachusetts Department of Environmental Protection including Gary Gonyea, Steve Pearlman, and Robert Nuzzo.

TABLE OF CONTENTS

FOREWORD

This paper describes a weight-of-evidence evaluation procedure for integrating the results of multiple measurements in environmental risk assessments. Multiple measurements are often used to evaluate each effect of concern. A weight-of-evidence evaluation takes into account the strengths and weaknesses of different measurement methods when determining whether the results show that a stressor has caused, or could cause, a harmful environmental effect.

The procedure outlined in this paper was developed by the Massachusetts Weight-of-Evidence Workgroup, an independent ad hoc group of ecological risk assessors from both government and the private sector. The Weight-of-Evidence Workgroup grew out of the Massachusetts Environmental Risk Characterization Guidance Workgroup, which has met intermittently since 1993 to assist the Massachusetts Department of Environmental Protection (DEP) in developing general guidance for risk characterization at disposal sites pursuant to the Massachusetts Contingency Plan (MCP). Workgroup members recognized that weight-of-evidence evaluation is a critical component in environmental risk assessments in general, and convened the Weight-of-Evidence Workgroup to focus on this topic.

The weight-of evidence project was conducted independently from the Massachusetts DEP Workgroup as these issues are broadly applicable, and because there was no previously published general guidance on the topic. It was hoped that operating outside of the constraints of a particular program would foster more creative thought, rigorous analysis and more open discussions in the workgroup. The Workgroup believes that a generally applicable evaluation method will provide a more solid foundation for further method development than would a program-specific methodology. Finally, Massachusetts DEP staff recognize that development of a broadly applicable method is an important first step in developing program specific guidance.

The Workgroup focused on developing a standard, *quantitative* evaluation procedure, which is described in detail in this paper. Most members of the Workgroup believe that a standard procedure will minimize subjectivity and promote consistency among assessments conducted by different risk assessors. (Although Massachusetts DEP staff participated in the Weight-of-evidence Workgroup, the Massachusetts Draft Environmental Risk Characterization Guidance currently recommends a *qualitative* approach that is based on the same criteria.)

A formal weight-of-evidence evaluation, whether qualitative or quantitative, can provide a framework for rigorous consideration of the strengths and weaknesses of various measurements, and of the nature of uncertainty associated with each of them. Applying a weight-of-evidence evaluation in an environmental risk assessment will promote systematic analysis by the risk assessor, and documentation of the evaluation will elucidate the risk assessor's thought process. It is important to recognize, however, that professional judgement may also be influenced by factors other than scientific knowledge and technical expertise.

Professional judgement applied in the selection and evaluation of measurements may incorporate both *knowledge* about the strengths and weaknesses of various measurements and *beliefs* about whether the measurements in question are likely to overestimate or underestimate risk. Thus, regulatory agency risk assessors, who are charged with *protection* against harm, may tend to be skeptical about the reliability of field studies, which provide direct measures of effects but may not have sufficient power to detect effects that could be biologically significant. At the same time, risk assessors representing the regulated community may be more wary of indirect non-site specific measurement methods, such as comparing contaminant concentrations to benchmark values published in the literature, which often suggest effects that are not observed in the field. A formal weight-of-evidence evaluation will not eliminate the influence of such beliefs from professional judgement. It may, though, increase risk assessor's awareness of his/her beliefs, and elucidate for the user/reviewer of the assessment the influence of beliefs on professional judgement.

Within the larger risk assessment community, the proposal outlined in this paper should not be viewed as a final product, but as a first step in the continuing effort to integrate diverse measurement methods in environmental risk assessments and to establish a framework for interpreting the results. The workgroup welcomes critical analysis of this proposal and encourages other efforts to the further develop guidance for weight-of-evidence evaluations.

EXECUTIVE SUMMARY

Weight-of-evidence is the process by which multiple measurement endpoints are related to an assessment endpoint to evaluate whether significant risk of harm is posed to the environment. In this paper, a methodology is offered for reconciling or balancing multiple lines of evidence pertaining to an assessment endpoint.

Weight-of-evidence is reflected in three characteristics of measurement endpoints: a) the weight assigned to each measurement endpoint; b) the magnitude of response observed in the measurement endpoint; and c) the concurrence among outcomes of multiple measurement endpoints. Briefly, the methodologies proposed to account for these three components are as follows.

First, weights are assigned to measurement endpoints based on attributes related to: a) strength of association between assessment and measurement endpoints; b) data quality; and c) study design and execution. These general categories are further divided into ten specific attributes. The relative importance of the ten specific attributes is then scaled, based on a survey of the professional judgement of ten ecological risk assessors. The resultant scaling values for the ten attributes range from 0.2 to 1.0. Measurement endpoints are then scored with respect to the ten attributes. Scores may range from one (low) to five (high). Unambiguous definitions of one through five for each attribute are provided, to limit subjectivity in the analysis. Finally, the weight of each measurement endpoint is obtained by multiplying the scaling values by the scores assigned for each attribute, summing these products, and dividing by 5.

Second, the magnitude of response in the measurement endpoint is evaluated with respect to whether the measurement endpoint indicates the presence or absence of harm (yes, no, or undetermined) and whether the response is low or high. In order to evaluate magnitude of response, the measurement endpoint is accompanied by a set of metrics, such as: a) change or difference in the response variable that is considered potentially ecologically relevant; b) spatial scale of the change or difference; and c) temporal scale of the change or difference.

Third, concurrence among measurement endpoints is evaluated by plotting the findings of the two preceding steps on a matrix for each measurement endpoint evaluated. The columns of the matrix present the weights assigned in the first step (e.g., 1-5), while the rows of the matrix present the magnitude of effect (e.g., positive effect of high magnitude, positive effect of low magnitude, undetermined, negative effect of low magnitude, negative effect of high magnitude). The matrix allows easy visual examination of agreements or divergences among measurement endpoints, facilitating interpretation of the collection of measurement endpoints with respect to the assessment endpoint.

For some risk assessments, the quantitative weight-of-evidence approach described above may be unwarranted, such as when measurement endpoints for a single assessment endpoint do not

contradict one another, or when a contradiction exists but there is a clear difference in the scientific defensibility of the endpoints. In these cases, the weight-of-evidence approach may be substantially simplified. A qualitative adaptation of the weight-of-evidence approach also involves three main steps; only the first step differs substantially from that applied under the quantitative method. First, each measurement endpoint is assigned a qualitative score of high, medium or low for each of the three principal attributes. The numbers of high, medium, and low scores for each measurement endpoint are counted and the measurement endpoint is assigned an overall score based on the majority of attribute specific scores. Second, the risk assessor evaluates the outcome of each measurement endpoint with respect to indication of risk of harm (e.g., positive, negative, or undetermined) and magnitude of the outcome (e.g., high or low). Third, the risk assessor integrates the measurement endpoint weight and magnitude of response on a matrix, in order to determine whether the overall evidence indicates a risk of harm. While this qualitative adaptation is clearly simpler to apply than the quantitative approach, it introduces greater subjectivity and may require less deliberate justification for conclusions regarding the potential risk of harm to the environment.

1.0 INTRODUCTION

A weight-of-evidence approach is recognized to be is a central component of ecological risk assessment. However, there is little specific guidance on the features of evaluating weight-of-evidence. An ad-hoc workgroup was formed to define what was meant by a weight-of-evidence approach and to outline a methodology for implementing such an approach. This group, comprised of representatives from the DEP, the U.S. Environmental Protection Agency (USEPA), the National Oceanic and Atmospheric Administration (NOAA), and six environmental consulting firms, met on an approximately monthly basis from November 1994 to June 1995. The group systematically examined various aspects of a weight-of-evidence approach and developed a method that reflects and makes transparent the underlying professional judgements associated with using a weight-of-evidence approach to characterize ecological risks. This paper presents the results of the workgroup meetings and proposes a method for implementing a weight-of-evidence approach. The ecological risk assessment terminology follows the USEPA's Framework for Ecological Risk Assessment (USEPA, 1992).

This paper is organized around the components of a weight-of-evidence approach, as defined below. The manner in which these components are incorporated into an ecological risk assessment are also described, along with a simple case example. Key points of discussion or analyses are highlighted throughout the paper.

1.1 Defining "Weight-of-Evidence"

Although the term "weight-of-evidence" is used frequently in ecological risk assessment, there is no consensus on its definition or how it should be applied. Published definitions or descriptions include:

> *"Each risk estimate will have its own assumptions and associated uncertainties and these may not be expressed equivalently. The separate lines of evidence must be evaluated, organized in some coherent fashion, and explained to the risk manager so that a weight-of-evidence evaluation can be made." Suter (1993).*
>
> *"Risk description has two primary elements. The first is the ecological risk summary, which summarizes the results of the risk estimation and uncertainty analysis and assesses confidence in the risk estimates through a discussion of the weight-of-evidence." EPA's Framework for Ecological Risk Assessment (USEPA, 1992).*
>
> *"For many Superfund ecological risk assessments, a weight-of-evidence approach will be used. This frequently will require that different types of data are evaluated together. These types of data may include toxicity test results, assessments of existing impacts on-site, or true risk calculations comparing estimated exposure doses with toxicity values from the literature. Balancing and interpreting the different types of data can be a major task...the*

strength of evidence provided by different types of tests and the precedence that one type of study has over another should already have been determined...This will insure that data interpretation is objective and not designed (i.e., biased) to support a preconceived answer." USEPA's Draft Ecological Risk Assessment Guidance for Superfund (USEPA, 1994).

The workgroup considered the available definitions and descriptions and derived a description that relates the weight-of-evidence approach to the process of conducting an ecological risk assessment:

> *"The weight-of-evidence approach is the process by which measurement endpoints are related to an assessment endpoint to evaluate whether a significant risk of harm[1] is posed to the environment. The approach is planned and initiated at the problem formulation stage and results are integrated at the risk characterization stage."*

This definition provides an explicit link between risk characterization and the assessment endpoints developed during problem formulation. Because the weight-of-evidence approach involves the process of relating measurement endpoints to an assessment endpoint, these two terms are defined below.

Assessment endpoints are explicit expressions of the actual environmental value that is to be protected. They reflect social and ecological priorities and are expressed in a manner that can be evaluated through an objective scientific process. They are most useful when they are expressed in terms of a specific receptor (species, habitat, system) and a function or quality that is to be maintained or protected.

Examples of clearly defined and ecologically relevant assessment endpoint are:

- maintenance of a benthic community that can serve as a prey base for local fish populations;
- reproductive success of the mink population within foraging range of the contaminated area; and
- community structure and reproductive success of songbird populations within a contaminated area.

Measurement Endpoints are the lines of evidence used to evaluate the assessment endpoint. Multiple measurement endpoints are often associated with a single assessment endpoint. The measurement endpoints are the bases for structuring the analysis phase of an ecological risk

[1] "Significant risk of harm" is the term used within the Massachusetts MCP to describe an unacceptable risk outcome.

assessment and serve as the actual measurements used to estimate risk. Therefore, they should be explicitly related - either directly or indirectly - to specific assessment endpoints. Further, they should include metrics (e.g., degree of response, space, and/or time) that can be used as a basis for estimating risks.

Examples of appropriate measurement endpoints for the example assessment endpoint, *maintenance of a benthic community that can serve as a prey base for local fish,* are:

- concentration of chemical of concern in sediment, relative to levels reported in the scientific literature to be harmful;
- toxicity observed in a whole sediment bioassay at levels considered significant according to the test protocol; and
- benthic invertebrate community structure, relative to reference areas.

1.2 Components of a Weight-of-Evidence Approach

The workgroup identified three major components that reflect the weight-of-evidence of measurement endpoints, with respect to a specific assessment endpoint:

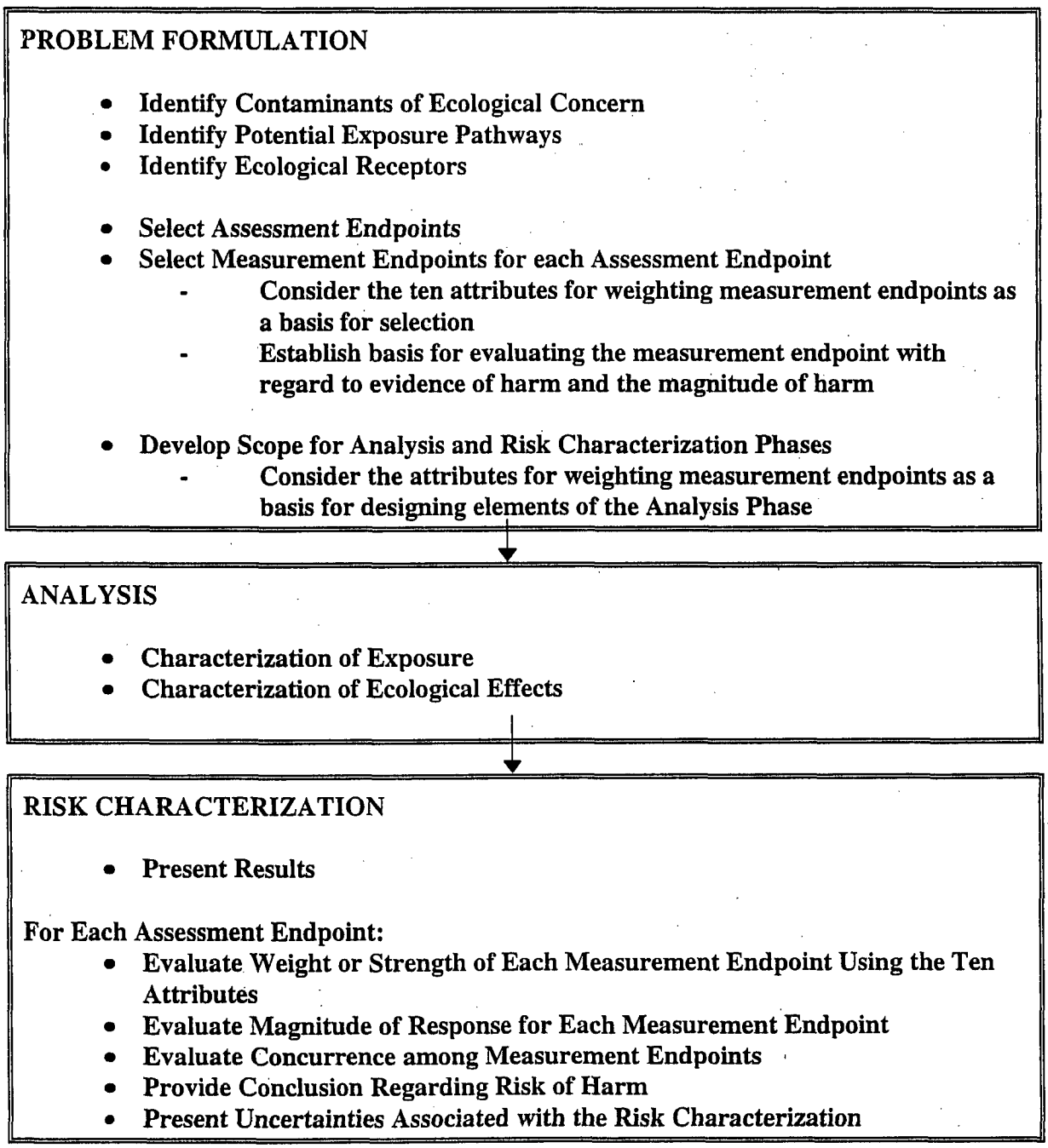
1. ***Weight assigned to each measurement endpoint***: Measurement endpoints may vary in the degree to which they relate to the assessment endpoint, the quality of the data, or the manner in which they were applied. Based on these attributes, an investigator may assign more weight to or have more confidence in one measurement endpoint compared to another.

2. ***Magnitude of response in the measurement endpoint***: Strong or obvious responses are typically assigned greater weight than marginal or ambiguous responses.

3. ***Concurrence among measurement endpoints***: More weight or confidence is generally attributed to findings in which there is agreement among multiple measurement endpoints. An investigator generally has less confidence in findings in which the lines of evidence contradict one another.

In the following sections, a weight-of-evidence approach is presented, based on these three components (Sections 2 - 4). The approach is quantitative inasmuch as numerical values are assigned to elements of professional judgement; the judgements themselves can be based on a number of qualitative as well as quantitative considerations. Section 5 presents a qualitative (i.e., non-numerical) approach to weighing evidence. The conceptual approach is illustrated in Figure 1.

Figure 1
Implementing a Weight-of-Evidence Approach
Within an Ecological Risk Assessment for Contaminated Sites
(Note: this is a simplified figure of the overall ecological risk process)



2.0 SELECTING AND WEIGHING MEASUREMENT ENDPOINTS

A weight-of-evidence evaluation enables the risk assessor to follow a process to determine on a relative scale those measurement endpoints that best represent the assessment endpoint, so that they have greater influence on the conclusions of the risk assessment. The approach outlined in this paper involves considering specific attributes of each measurement endpoint to determine how well the endpoint represents the assessment endpoint. Attributes are defined as the characteristics of a measurement endpoint that determine how well it estimates or predicts the effect defined by the assessment endpoint. Consideration of the specific attributes defined later in this paper enable the risk assessor to identify the measurement endpoints that best represent the assessment endpoints, and to give them more weight in the risk assessment.

> Some members of the group felt that the attribute related to quality of data should be evaluated separately and prior to the weighing of the measurement endpoints. As a result, weighing schemes were developed with and without his attribute.

The workgroup identified ten attributes considered most important in selecting and weighing measurement endpoints. These were grouped into three categories: (1) attributes related to strength of association between assessment and measurement endpoints; (2) attributes related to data quality and (3) attributes related to study design and execution. Table 1 presents definitions of the attributes in each of these groups.

2.1 Application of Measurement Endpoint Attributes in the Assessment Process

The question of how well a measurement endpoint represents an assessment endpoint arises at two separate points in the risk assessment process: (1) in the problem formulation stage, when the risk assessor selects optimal measurement endpoints for evaluating each assessment endpoint, and (2) in the risk characterization stage, when the risk assessor evaluates whether the results of various measurements (which may contradict each other) indicate a risk of harm to the environment. Thus, the ten attributes listed in Table 1 are applied both in selecting measurement methods (and endpoints) and in weighing different results obtained from different measurements. Consideration of these attributes during selection of measurement endpoints in the problem formulation phase will help ensure that the overall study design is strong and scientifically defensible and that the findings of the ecological risk assessment are conclusive.

Table 1

Attributes for Judging Measurement Endpoints

In the weight-of-evidence procedure, ten attributes of each measurement endpoint are evaluated. For a given assessment endpoint, the quality of each measurement endpoint is compared with respect to these attributes. Those measurement endpoints with the highest quality for the most attributes are given the greatest weight in the overall characterization of risk. The attributes for consideration are:

I. Attributes Related to Strength of Association Between Assessment and Measurement Endpoints

Biological linkage between measurement endpoint and assessment endpoint.
This attribute refers to the extent to which the measurement endpoint is representative of, and correlated with, or applicable to the assessment endpoint. If there is no biological linkage between a measurement endpoint (e.g., a study that may have been performed for some other purpose) and the assessment endpoint of interest, then that study should not be used to evaluate the stated assessment endpoint. Biological linkage pertains to similarity of effect, target organ, mechanism of action, and level of ecological organization.

Correlation of stressor to response.
This attribute relates the ability of the endpoint to demonstrate effect from chronic exposure to the stressor and to correlate effects with the degree of exposure. As such, this attribute also takes into consideration the susceptibility of the receptor and the magnitude of effects observed.

Utility of measure for judging environmental harm.
This attribute relates the ability to judge results of the study against well-accepted standards, criteria, or objective measures. As such, the attribute describes the applicability, certainty, and scientific basis of the measure, as well as the sensitivity of a benchmark in detecting environmental harm. Examples of objective standards or measure for judgment might include ambient water quality criteria, sediment quality criteria, biological indices, and toxicity or exposure thresholds recognized by the scientific or regulatory community as measures of environmental harm.

II. Attributes Related to Data Quality

Extent to which data Quality Objectives are Met.
This attribute reflects to degree to which data quality objectives are designated that are comprehensive and rigorous, as well as the extent to which they are met. Data quality objectives should clearly evaluate the appropriateness of data collection and analysis practices. If any data quality objectives are not met, the reason for not meeting them and the potential impact on the overall assessment should be clearly documented.

III. Attributes Related to Study Design and Execution

Site-specificity.
This attribute relates the extent to which chemical and biological data, environmental media, specie, environmental conditions, benchmark (or reference), and habitat types that are used in the measurement endpoint reflect the site of interest.

Sensitivity of the measurement endpoint for detecting changes.
This attribute relates to the ability to detect a response in the measurement endpoint, expressed as the percentage of the total possible variability that the endpoint is able to detect. Additionally, this attribute reflects the ability of the measurement endpoint to discriminate between responses to a stressor and those resulting from natural or design variability and uncertainty.

Spatial representativeness.
This attribute relates to the degree of compatibility or overlap between the study area. locations of measurements or samples, locations of stressors, and locations of ecological receptors and their potential exposure.

Temporal representativeness
This attribute relates to the temporal compatibility or overlap between the measurement endpoint (when data were collected or the period for which data are representative) and the period during which effects of concern would be likely to be detected. Also linked to this attribute is the number of measurement or sampling events over time and the expected variability over time.

Quantitativeness
This attribute relates to the degree to which numbers can be used to describe the magnitude of response of the measurement endpoint to the stressor, as well as whether results are objective or subjective, whether the results are sufficient to test for statistical significance, and whether biological significance can be inferred from statistical significance.

Use of a standard method.
The extent to which the study follows standard protocols recommended by a recognized scientific authority for conducting the method correctly. Examples of standard methods are study designs or chemical measures published in the Federal Register of the Code of Federal Regulations, developed by ASTM, or repeatedly published in the peer reviewed scientific literature, including impact assessments, filed surveys, toxicity tests, benchmark approaches, toxicity quotients, and tissue residue analyses. This attribute also reflects the suitability and applicability of the method to the endpoint and the site, as well as the need for modification of the method.

Selecting and linking measurement endpoints to assessment endpoints provides a basis for planning the scope of the Analysis phases of the risk assessment.

Consideration of these attributes in the risk characterization phase fosters a systematic and balanced consideration of the strengths and weaknesses of the information derived from each measurement approach. Further, a full discussion of how the attributes are considered in the weight-of-evidence evaluation elucidates the risk assessor's thought process and professional judgements. A rigorous explanation of the links and gaps between the measurement results and the risk assessor's conclusions enables risk mangers to make decisions with a clear understanding of the uncertainties inherent in the assessment.

Some members of the workgroup felt that, when using a weight-of-evidence evaluation to characterize risk, the attribute related to quality of data should be evaluated separately and prior to the weighing of the measurement endpoint. One suggested approach is to consider data quality as a pass/fail criterion, as follows:

- if the quality is adequate, the measurement endpoint is retained for consideration in the risk characterization, but the data quality is not considered as a factor in the weight-of-evidence evaluation;

- if the data quality is inadequate, the endpoint is not considered in the risk characterization step.

Other members of the workgroup felt that data quality should be fully considered in the weight-of-evidence evaluation. As a result, weighing schemes were developed both with and without the attribute.

A weight-of-evidence evaluation may be qualitative or quantitative. The workgroup focused on a developing a quantitative approach, because many members felt that a quantitative scheme would be applied more consistently and would minimize subjectivity. However, a qualitative evaluation would use the same attributes. Quantitative and qualitative approaches are discussed in the following sections.

2.2 Consideration of Attributes in a Qualitative Weight-of-evidence Evaluation

When selecting measurement methods or characterizing risk, the attributes listed in Table 1 can be considered qualitatively, without assigning any numerical values. The evaluations should consider both the *relative importance of each attribute* and the *quality (or efficacy) of the measurement endpoint with respect to each attribute.*

For different assessment and measurement endpoints, the relative importance of some of the attributes may vary. One advantage of a qualitative approach is that the relative importance of each attribute is not fixed, and can be considered differently on a case by case basis.

In a qualitative weight-of-evidence evaluation, the efficacy of the measurement endpoint relative to each attribute can be described in non-numeric terms. The qualitative approach described in Section 5.0 uses ratings of high, medium and low to describe the quality a measurement endpoint with respect to individual attributes and groups of attributes.

2.3 Consideration of Attributes in a Quantitative Weight-of-evidence Evaluation

In a quantitative weight-of-evidence evaluation, the attributes are used to assign weights to each measurement endpoint. The process of assigning weights to measurement endpoints incorporates two elements:

1. The relative importance of each attribute: Investigators consider some attributes more important than others when considering the overall weight of measurement endpoints.

2. The scores that a measurement receives with respect to each attribute: When measurement endpoints are compared with respect to each of the attributes, some will score better than others.

These two elements of assigning weights to measurement endpoints are described in the next three subsections.

2.3.1 Scaling the Relative Importance of Attributes

The ten attributes can either be assigned equal importance or they can be scaled to reflect their relative importance in weighing measurement endpoints. The relative importance of each attribute is subjective and reflects professional judgement. To facilitate implementation of the weight-of-evidence approach, the workgroup developed a set of fixed scaling values that reflect collective professional judgement and can be

There was considerable discussion within the group regarding the merits of scaling the attributes to reflect relative importance. A strong case was made for treating them all equally. However, the group decided to proceed to develop a scaling system.

applied to ecological risk assessments.

If an investigator chooses to diverge from this fixed set of scaling values, he or she can present an alternative set of scaling values and rationale for their use. However, a set of values based on collective professional judgement reflects the range of opinion that exists among scientists; as such, bias that may be held by any one scientist, is minimized or avoided altogether. The set of scaling values described below were developed based on a survey of ten experienced ecological risk assessors. Because of variability among individuals' professional judgement a survey of a larger or different group of ecological risk assessors might yield somewhat different scaling values.

Survey of Ecological Risk Assessors

Ten ecological risk assessors participated in a survey with the objective of scaling the relative importance of the ten attributes listed in Table 1. The participants were provided with a matrix listing the attributes horizontally and vertically. They were provided with the following instructions:

1. Score the attributes on the top (horizontal) row of the diagram against the ones on the left (vertical) column using "+s", "-s", or 0 as follows:

   +++ if the attribute in the top row is much more important than the attribute in the left column;

   ++ if the attribute in the top row is more important than the attribute in the left column;

   + if the attribute in the top row is slightly more important than the attribute in the left column;

   0 if the attribute in the top row is as important as the attribute in the left column;

   - if the attribute in the top row is slightly less important than the attribute in the left column;

   -- if the attribute in the top row is less important than the attribute in the left column;

   --- if the attribute in the top row is much less important than the attribute in the left column.

2. Please answer the following questions:
Which attribute(s) do you think are most important on a relative basis?:
Which attribute(s) do you think are least important on a relative basis?:
Please check how much greater importance would you give to the most important attribute(s) as compared to the least important ones?:

they're all of equal importance ____
2 x as important ____
5 x as important ____
10 x as important ____
20 x as important ____
50 x as important ____
100 x as important _____

The "+s" and "-s" in each participant's matrix were converted to numerical values ranging from 3 (for 3 +s) to -3 (for 3 -s). Each combination of values was entered into a table and the average and range of values was obtained for each pair of attributes. The average and ranges provided an indication of the relative importance of each of the attributes.

As shown in the second section of the survey, participants also provided information on how much more important they viewed the most important attribute relative to the least. Respondents gave values that ranged from 2 to 50. The geometric mean of the values was 11. The survey results indicated that people differed slightly on their choices of the most and least important attributes. Therefore, the range was adjusted to reflect the average spread in the following manner. Using the "+" and "-" system, the maximum spread between any two attributes was 3 (i.e., 3 +s or 3 -s) but the maximum spread among the averaged values was 1.8, or 60% of the total possible range. The geometric mean range of 11 was multiplied by 0.6 to yield an adjusted range of 6.6. The most important attribute was assigned a scaling value of 1.0 and values for other attributes were adjusted to correspond with their relative importance and to fit within an overall range of 6.6. The results were then rounded to one significant figure which yielded an overall range of 5 (from 0.2 to 1.0). Because the workgroup could not reach consensus on the most appropriate role of the attribute data quality in the overall weight-of-evidence approach, scaling values were calculated that both included and excluded the attribute related to the quality of data. The resultant values for scaling the relative importance of the attributes are given in Table 2. The values provided in the table could be applied to most ecological risk assessments, inasmuch as they reflect collective judgement independent of the measurement endpoints that they are used to help weigh.

Table 2
Value Scale Representing the Relative Importance of Attributes

Attribute	Scaling Values Including Quality of Data	Scaling Values Excluding Quality of Data
Degree of Association	1.0	1.0
Stressor/Response	0.7	0.6
Utility of Measure	0.5	0.4
Quality of Data	0.8	X
Site Specificity	0.5	0.5
Sensitivity	0.5	0.5
Spatial Representativeness	0.4	0.4
Temporal Representativeness	0.2	0.2
Quantitative Measure	0.2	0.2
Standard Measure	0.2	0.2

2.3.2 Scoring the Attributes

When evaluating measurement endpoints using the ten attributes it can be expected that the endpoints will conform with the attributes to varying degrees. The workgroup developed guidelines for scoring a measurement endpoint against each attribute to quantify this variability. A range in score from one (low) to five (high) was selected, because it was perceived as having a broad enough spread to allow differentiation between scores for measurement endpoints, without being overly cumbersome. The workgroup established non-overlapping, comprehensive, and broadly applicable criteria based on the most relevant considerations for each attribute for assigning numeric scores to measurement endpoints (Table 3).

Table 3

Definition of Scores Applied to Endpoint-Attribute Pairs in Weight of Evidence for Ecological Risk Assessment
I. Attributes Related to Strength of Association Between Assessment and Measurement Endpoints

Attribute	Factors to Consider in Ranking	1	2	3	4	5
Biological linkage between measurement endpoint and assessment endpoint	Correlation and/or applicability of measurement endpoint with respect to assessment endpoint; linkage based on known biological processes; similarity of effect, target organ, mechanism of action, ad level of ecological organization	Biological processes link the measurement endpoint to the assessment endpoint only indirectly*, yielding a weak correlation between the assessment and measurement endpoints	Biological process directly links the measurement and assessment endpoints, although the specific effect, target organ, and mechanism of action evaluated are not the same	Measurement and assessment endpoints are directly linked and the adverse effect, target organ, and mechanism of action are the same for both endpoints; however, the levels of ecological organization differ**	Measurement and assessment endpoints are directly linked and the adverse effect, target organ, mechanism of action, and level of ecological organization are the same for both endpoints	Assessment endpoint is directly measured and, therefore, is equivalent to the measurement endpoint.
Correlation of stressor to response	Ability of endpoint to demonstrate effects from chronic exposure to stressor and to correlate effects with degree of exposure; susceptibility and magnitude of effects.	Endpoint response to stressor has not been demonstrated in previous studies but is expected to, based upon demonstrated response to similar stressors	In previous studies, endpoint response to stressor has been suggested, but has not been definitely proven	In previous studies, endpoint response to stressor has been demonstrated, but response is not correlated with magnitude of exposure	Response is quantitatively correlated with magnitude of exposure, but correlation is not statistically significant (or data are not sufficient to test for statistical significance)	Statistically significant correlation is demonstrated
Utility of measure for judging environmental harm	Applicability, certainty, and scientific basis of measure that is used to judge environmental harm; sensitivity of benchmark in detecting environmental harm	Measure is developed by the investigator (i.e., personal index) and has limited applicability and certainty and the scientific basis is weak and the benchmark is relatively insensitive	Measure is personal index and has either limited applicability or certainty or the scientific basis is weak or the benchmark is relatively insensitive	Measure is well accepted and developed by a third party but has either limited applicability or certainty or the scientific basis is weak or the benchmark is relatively insensitive	Measure is well accepted and developed by a third party and has moderate certainty, applicability and scientific basis and benchmark is moderately sensitive	Measure is well accepted and developed by a third party and has very high levels of certainty and applicability, as well as a very strong scientific basis and benchmark is very sensitive

* An example of an indirect biological link is measurement of community structure for the assessment endpoint of neurotoxicity
** An example of differing levels of ecological organization is measurement of impacts to individual organisms of a single species

Table 3 (continued)

Definition of Scores Applied to Endpoint-Attribute Pairs in Weight of Evidence for Ecological Risk Assessment
III. Attributes Related to Study Design and Execution

Attribute	Factors to Consider in Ranking	1	2	3	4	5
Site-specificity	Representativeness of chemical or biological data, environmental media, species, *environmental conditions*, benchmark (or reference) and habitat types that are used in the measurement endpoint relative to those present at the site	Only one or two of the six factors (i.e., data, media, species, env. conditions, benchmark, habitat type) is derived from or reflects the site	Three of the six factors are derived from or reflect the site	Four of the six factors are derived from or reflect the site	Five of the six factors are derived from or reflect the site	All six factors (i.e., data, media, species, env. conditions, benchmark, habitat type) are derived from or reflect the site (i.e., both data and benchmark reflect site conditions)
Sensitivity of the *measurement* endpoint for detecting changes	The percentage of the total possible variability that the endpoint is able to detect; ability of measurement endpoint to detect effects from stressor, rather than from natural or design variability or uncertainty	Endpoint can detect changes larger than 1,000X	Endpoint can detect changes between 100X and 1,000X	Endpoint can detect changes between 10X and 99X	Endpoint can detect changes between 2X and 9X	Endpoint can detect changes of less than 2X
Spatial represent-ativeness	Spatial overlap of study area, measurement or sampling stations, locations of stressors, locations of receptors, and points of potential exposure to those receptors*	The locations of two of the following subjects overlap spatially only to a limited extent: study area, sampling/measurement site, stressors, receptors, and points of potential exposure	The locations of two of the following subjects overlap spatially: study area, sampling/measurement site, stressors, receptors, and points of potential exposure	The locations of three of the following subjects overlap spatially: study area, sampling/measurement site, stressors, receptors, and points of potential exposure	The locations of four of the following subjects overlap spatially: study area, sampling/measurement site, stressors, receptors, and points of potential exposure	The locations of five of the following subjects overlap spatially: study area, sampling/measurement site, stressors, receptors, and points of potential exposure
Temporal *representativeness*	Temporal overlap between the *measurement period* and the period during which chronic effects would be likely to be detected (daily, weekly, seasonally, annually),	Measurements are collected during a season different from when effects would be expected to be most clearly manifested; AND	Measurements are collected during a season different from when effects would be expected to be most clearly manifested; OR	Measurements are collected during the same period that effects would be expected to be most clearly manifested; AND	Measurements are collected during the same period that effects would be expected to be most clearly manifested; AND	Measurements are collected during the same period that effects would be expected to be most clearly manifested; AND

TABLE 3 (continued)

Definition of Scores Applied to Endpoint-Attribute Pairs in Weight of Evidence for Ecological Risk Assessment
II. Attributes Related to Data Quality

Attribute	Factors to Consider in Ranking	1	2	3	4	5
Quality of data	Extent to which DQOs* are met	Three or more DQOs are not met OR	Two DQOs are not met AND	One DQO is not met AND	One DQO is not met and	All DQOs are met AND
		DQOs barely meet the needs of the risk assessment OR	DQOs meet the needs of the risk assessment satisfactorily AND	DQOs meet the needs of the risk assessment satisfactorily AND	DQOs are rigorous and comprehensive AND	DQOs are rigorous and comprehensive
		There is no documentation of the reason for not meeting DQO and the impact on the assessment	Reason for not meeting DQOs and the impact on the assessment are documented satisfactorily	Reason for not meeting DQO and the impact on the assessment are clearly documented	Reason for not meeting DQO and the impact on the assessment is clearly documented	

Note:

A field and Laboratory Reference. EPA600 3-89/013.

Table 3 (continued)

Definition of Scores Applied to Endpoint-Attribute Pairs in Weight of Evidence for Ecological Risk Assessment
III. Attributes Related to Study Design and Execution

Attribute	Factors to Consider in Ranking	1	2	3	4	5
Temporal representativeness (continued)	Number of measurement or sampling events over time, and Expected variability over time	A single sampling or measurement event is conducted; AND High variability in that parameter is expected over time	[A single sampling or measurement event is conducted; AND High variability in that parameter is expected over time]	A single sampling or measurement event is conducted; AND Moderate variability in that parameter is expected over time.	Two sampling or measurement events are conducted; AND Moderate variability in that parameter is expected over time	EITHER [two sampling events are conducted and variability is low OR multiple sampling events are conducted and variability is moderate to high]
Quantitativeness	Results are quantitative/qualitative, subjective/objective, sufficient to test for statistical significance, and extent to which biological significance	Results are qualitative and are subject to individual interpretation	Results are qualitative and are not subject to individual interpretation (i.e., objective)	Results are quantitative, but data are insufficient to test for statistical significance	Results are quantitative and may be tested for statistical significance, but such tests do not clearly reflect biological significance	Results are quantitative and may be tested for statistical significance; such tests clearly reflect biological significance
Use of a standard method	Method availability; ASTM approval; suitability & applicability to endpoint and site; need for modification of method; relationship to impact assessment, field survey, toxicity test, benchmark, toxicity quotient, or tissue residue analysis methodologies	Method has never been published AND methodology is not an impact assessment, field survey, toxicity test, benchmark approach, toxicity quotient, or tissue residue analysis	Method is one of the 6 listed methodologies, but the particular application is neither published nor standardized	A standard method exists, but its suitability for this purpose is questionable, and it must be modified to be applicable to site specific conditions	A standard method exists and it is directly applicable to the measurement endpoint, but it was not developed precisely for this purpose and requires slight modification OR the methodology is used in two peer-reviewed studies	A standard method exists and is directly applicable to the measurement endpoint and it was developed precisely for this purpose and requires no modification OR the methodology is used in three or more peer-reviewed studies

* Study area, sampling station, and points of exposure are differentiated by level of specificity. While the study area may be a 5 acre wetland, sampling stations may o the 2 acres that are accessible, while the actual points of exposure to invertebrate receptors may be the top 6 inches of sediment

2.4 Weighing Measurement Endpoints

The weight of a measurement endpoint is obtained by multiplying the scaling values (Table 2) by the scores the measurement endpoint is assigned for each attribute (using Table 3), summing the products for each measurement endpoint, and dividing by 5 (or 4 if quality of data is excluded), to yield weighing values that range between 1 and 5:

Measurement endpoint weight = Σ (scaling value * score)/5

The measurement endpoint weights are then rounded to the nearest whole number. Spreadsheets can be used to automate the calculations, as illustrated in Table 4. This step provides a quantitative measure of the first component of the weight-of-evidence - the weight given each measurement endpoint. The workgroup discussed the most appropriate number of significant figures for the weights of individual measurement endpoints, but did not reach a consensus. Whole numbers are a simplification and reflect the limited precision of the process However, two or more significant figures enable the risk assessor to more clearly differentiate between two measurement endpoints with similar weights. Additional case studies may elucidate which approach is more appropriate.

Example:
Determining Weight of Measurement Endpoints

To illustrate the weight-of-evidence approach, the following example is used. Sediments of a river have been contaminated with an organic chemical that is acutely and chronically toxic to aquatic life. Several assessment endpoints have been developed for evaluating risks at the site. One of the assessment endpoints is *maintenance of a benthic community that can serve as a prey base for local fish.* Three measurement endpoints were chosen to evaluate the assessment endpoint: A) the concentration of chemical in the sediments in relation to levels reported to be harmful; B) toxicity as measured in a whole sediment bioassay, where mortality in excess of 20% is considered an adverse effect; and C) abundance and community structure of invertebrates that compose the diet of local fish species at and near the release location, as compared to reference areas. The risk assessor has examined these three measurement endpoints against the ten attributes and scored them in Table 5. The scores reflect a number of site-specific factors.

Table 4

Scoring Measurement Endpoints (Scheme A)

Score Each Measurement Endpoint from Low to High (1 - 5)

Assessment Endpoint: _____

Attributes Weighing Factors	Weighing Factors	Measurement Endpoint A	Measurement Endpoint B	Measurement Endpoint C
I Relationship Between Measurements and Assessment Endpoints				
• Degree of Association	1.0			
• Stressor/Response	0.7			
• Utility of Measure	0.5			
II Data Quality				
• Quality of Data	0.8			
III Study Design				
• Site Specificity	0.5			
• Sensitivity	0.5			
• Spatial Representativeness	0.4			
• Temporal Representativeness	0.2			
• Quantitative Measure	0.2			
• Standard Method	0.2			
(Sum scores*weighting factors)/5	1			
Round to nearest whole number				

3.0 MAGNITUDE OF RESPONSE IN THE MEASUREMENT ENDPOINT

As discussed in Section 1.2, the magnitude of the response in the measurement endpoint is considered together with the measurement endpoint weight in judging the overall weight-of-evidence. The workgroup divided magnitude of response into two questions:

1. Does the measurement endpoint indicate the presence or absence of harm (yes, no, or undetermined)?

2. Is the response low or high?

While these issues are presented above as discrete functions, the workgroup recognizes that responses are more likely to occur as continuous gradients, and that the risk assessment may present the results as such. However, the workgroup agreed that discrete categories accompanying a detailed analysis would more clearly communicate results to risk managers and others.

Metrics

In order to evaluate magnitude of response, the measurement endpoints must be accompanied by a set of metrics. Ideally, such metrics are established during the problem formulation stage, through discussions with the risk manager. They may be accompanied by a statement of the value considered statistically significant, if possible. In general, one or more of the following metrics is included for evaluating the response in the measurement endpoint:

1. A change or difference in the response variable that is considered potentially ecologically relevant (e.g., percent of mortality or change in abundance or biomass);

2. Spatial scale of the change or difference, as related to the assessment endpoint (e.g., hectares, fraction of foraging area, fraction of area utilized by a local population);

3. Temporal scale of the change or difference, as related to the assessment endpoint [duration, changes over time with and without natural stressors (e.g., as storms or floods), rate of recovery].

Prior to determining the magnitude of effect, the risk assessor should consider at what level(s) a response would be considered indicative of environmental harm with respect to the assessment endpoint. If possible, it is helpful to set specific criteria for establishing these thresholds. The risk assessor should consider, *a priori* what represents a "low" or "high" response along a response gradient. Within the analysis and risk characterization sections of a report, the risk assessor should present and discuss the details of the considerations and their interpretation.

Table 5

Scoring Measurement Endpoints (Scheme A)

Score Each Measurement Endpoint from Low to High (1 - 5)

Assessment Endpoint: ***Maintenance of a benthic community that can serve as a prey base for local fish.***

Attributes Weighing Factors	Weighing Factors	Measurement Endpoint A	Measurement Endpoint B	Measurement Endpoint C
I Relationship Between Measurements and Assessment Endpoints				
• Degree of Association	1.0	1	3	5
• Stressor/Response	0.7	3	5	2
• Utility of Measure	0.5	4	4	2
II Data Quality				
• Quality of Data	0.8	4	5	2
III Study Design				
• Site Specificity	0.5	3	5	5
• Sensitivity	0.5	4	4	2
• Spatial Representativeness	0.4	4	4	4
• Temporal Representativeness	0.2	3	3	3
• Quantitative Measure	0.2	4	4	4
• Standard Method	0.2	4	5	3
(Sum scores*weighting factors)/5	1	2.77	4.20	3.42
Round to nearest whole number		3	4	3

The weighting scores (e.g. 1-5), evidence of harm, and magnitudes of response are integrated for each measurement endpoint in a matrix such as that presented in Table 6. This summary table provides a simple communication tool and indicates the risk assessor's conclusions regarding the magnitude of response.

Example
Determining the Degree of Response in Measurement Endpoints

In the example of the assessment endpoint *maintenance of a benthic community that can serve as a prey base for local fish,* the risk assessor made the following determinations for each measurement endpoint, as illustrated in Table 7.

Measurement Endpoint A - the concentration of chemical in the sediments in relation to levels reported to be harmful indicated a low risk of harm

Measurement Endpoint B - toxicity as measured in a whole sediment bioassay indicated a high risk of harm

Measurement Endpoint C - abundance and community structure of invertebrates was undetermined (i.e., the design of the study and/or natural variability precluded a determination of either harm or lack of harm)

TABLE 6

Risk Assessment Scoring Sheet For

Evidence of Harm and Magnitude

Assessment Endpoint: _____

_____

Measurement Endpoints	Weighting Score (1 - 5)	Evidence of Harm (Yes/No/Undetermined)	Magnitude (High/Low)
Endpoint A			
Endpoint B			
Endpoint C			

TABLE 7. Examples of Risk Assessment Scoring Sheet For Evidence of Harm and Magnitude

Assessment Endpoint: Maintenance of a benthic community that can serve as prey base for local fish

Measurement Endpoints	Weighting Score (1 - 5)	Evidence of Harm (Yes/No/Undetermined)	Magnitude (High/Low)
Endpoint A	2.7 (3)	Yes	Low
Endpoint B	4.2 (4)	Yes	High
Endpoint C	3.4 (3)	Undetermined	Undetermined

4.0 CONCURRENCE AMONG MEASUREMENT ENDPOINTS

The third component of the weight-of-evidence approach involves examining concurrence among measurement endpoints as they relate to a specific assessment endpoint. Logical connections, interdependence, and correlations among measurement endpoints should also be considered when evaluating concurrence.

The workgroup developed a graphical method for displaying concurrence among measurement endpoints (Table 8). The method involves plotting the letter designation of the measurement endpoint within a matrix with weight of the measurement endpoint and degree of response as axes. The graphical method permits easy visual examination of agreements or divergences among measurement endpoints, along with the weights assigned to the endpoints.

Example
Examining Concurrence Among Measurement Endpoints

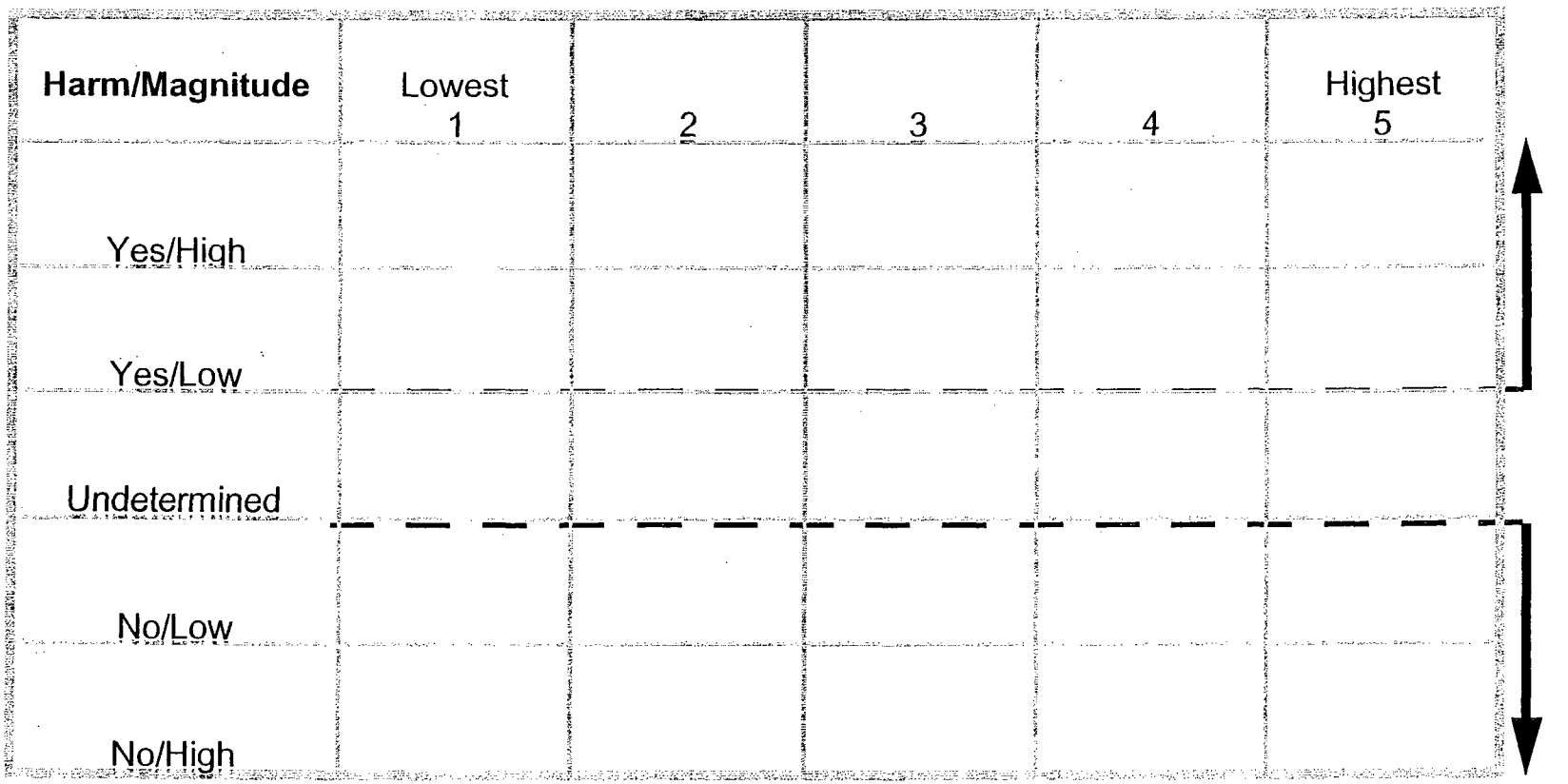
The letters associated with each of the three measurement endpoints used to evaluate *maintenance of a benthic community that can serve as a prey base for local fish* were plotted on the matrix (Table 9). The resulting plot shows that two of the three measurement endpoints indicated some risk of harm while the third was undetermined. The illustration also shows that Measurement Endpoint C (a field study) was assigned a low weight, even though it was a direct measure of the assessment endpoint. As shown in Table 5, the low weight reflected relatively poor design and data quality. The risk assessor and risk manager might reach the following conclusions based on Table 9: 1) there is a risk of harm to the environment as indicated by the preponderance of evidence and the relative weights of the measurement endpoints; and 2) the "undetermined" status of Measurement Endpoint C diminishes the overall conclusion that there is a risk of harm. Either the uncertainty associated with Measurement Endpoint C could be accepted or additional work could be conducted to strengthen the analysis.

Table 8
Risk Analyses Summary Sheet

Assessment Endpoint:

Weighing Factors

increasing confidence or weight



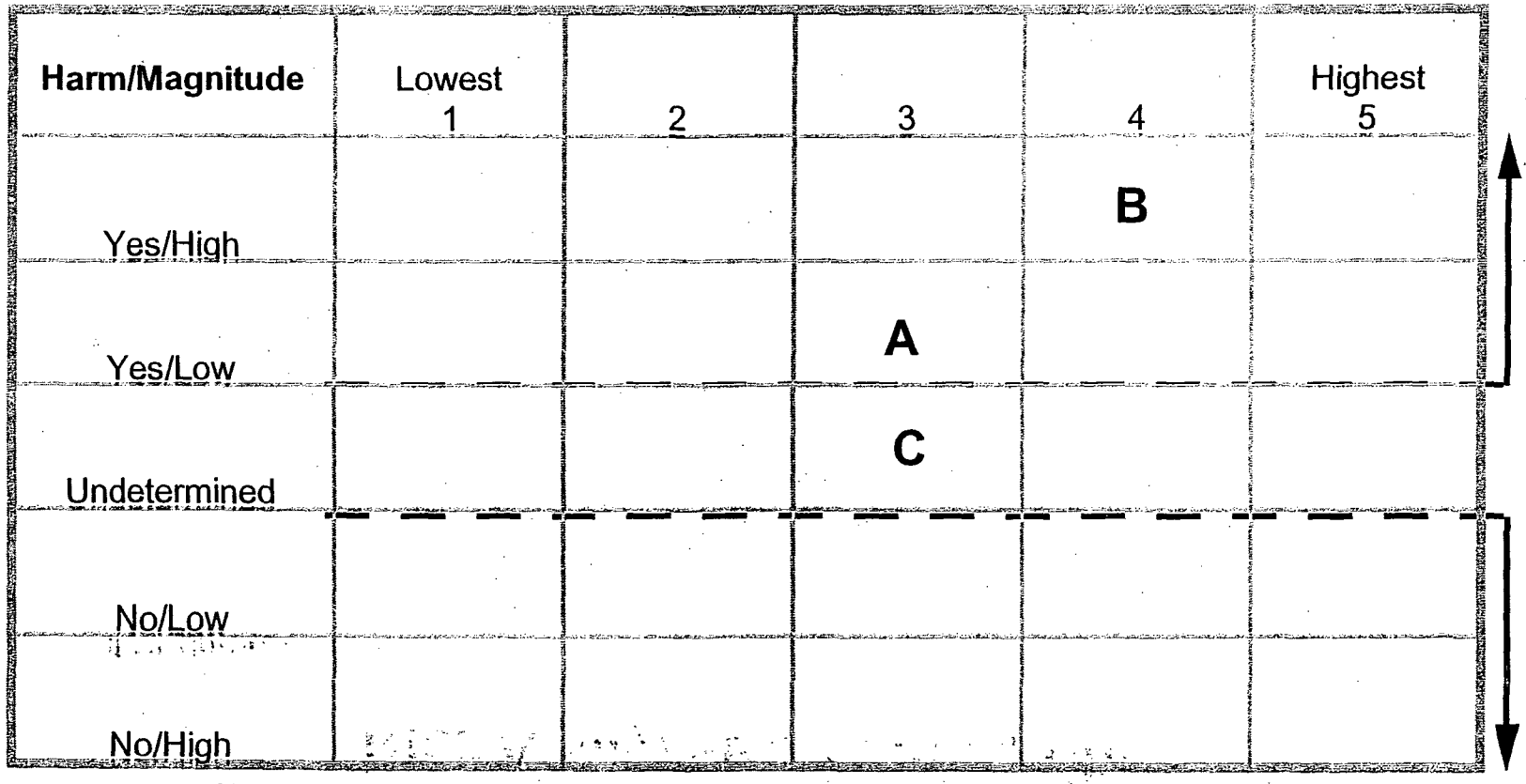
Use letter designations to place measurement endpoints in the boxes

Table 9
Example Risk Analyses Summary Sheet

Assessment Endpoint: **Maintenance of a benthic community that can serves as a prey base for local fish**

Weighing Factors

increasing confidence or weight



Use letter designations to place measurement endpoints in the boxes

5.0 QUALITATIVE WEIGHT-OF-EVIDENCE APPROACH

While most risk assessors and risk managers likely agree on the utility of applying a weight- of-evidence approach to ecological risk assessment, a quantitative method such as that described above may be perceived as inflexible or overly complicated for certain risk assessments. If desired, the approach may be adapted to be more qualitative, while still maintaining the process of characterizing professional judgements according to the attributes defined in this weight-of-evidence approach.

The qualitative adaptation of the weight of evidence approach consists of three main steps which parallel the components of the quantitative approach.

(1) Each measurement endpoint is assigned a score of high, medium or low for each of the ten individual attributes. Based upon those scores and on the relative importance of individual attributes, the risk assessor should determine an overall score of high, medium or low indicating how well the measurement endpoint represents the assessment endpoint.

If all attributes are assumed to be of equal importance, then scoring is a simple matter of counting high, medium and low scores. However, most risk assessors are likely to consider some attributes more important than others when assigning an overall score. Determining the relative importance of attributes is a subjective process involving professional judgement. Therefore, it is imperative that the risk assessor provide a detailed and comprehensive description of his\her decision process in order to make the conclusions meaningful to the risk manager.

The risk assessor may amend the matrix of score definitions (Table 3) to reflect the three qualitative categories, rather than the five quantitative categories. However, the attribute definitions and the criteria used to score the measurement endpoint with respect to each attribute must be clearly stated and fully explained. To insure a systematic evaluation and an unambiguous assessment documentation, the attributes must be clearly and rigorously defined by the risk assessor.

(2) The outcome of each measurement endpoint is evaluated with respect to magnitude of response. The indication of risk of harm to the environment is described as positive, negative, or undetermined indication of risk.

The magnitude of the outcome is determined, based on the definitiveness of a positive or negative result. The magnitude of the outcome may be characterized as high or low.

(3) Finally, the risk assessor integrates the measurement endpoint weight, and magnitude of response to determine whether the overall evidence indicates a risk of harm. To that end, each measurement endpoint (e.g., A, B, C) is placed on a matrix comparable to Table 10.

Table 10

Table for Integrating Overall Weight-of-evidence

HARM/ MAGNITUDE	LOW WEIGHT	MEDIUM WEIGHT	HIGH WEIGHT
Yes/High			A
Yes/Low		B	
Undeterminate		*	
No/Low	C		
No/High			

To assess the overall weight-of-evidence, the risk assessor may view Table 10 as a plane, the dot at the center of the intersection of "medium weight" and "undetermined" as a fulcrum, and the direction of tilt of the plane as the weight-of-evidence. In the example provided above, the plane would be tilted toward risk, since Endpoints A and B counterbalance Endpoint C, which has been assigned a low weight and yields only a weak indication that there is no risk.

In short, the main differences in methodology between the quantitative approach described in Sections 2 through 4 and the qualitative approach presented in this section pertain to (1) weighing the attributes and (2) scoring the measurement endpoints. Whereas the quantitative approach assigns fixed numerical weights to the ten attributes to reflect differing degrees of importance, the qualitative approach does not involve pre-assigned weights. The quantitative approach allows the risk assessor to use the scaling values to derive numerical scores for each assessment endpoint. The qualitative approach requires the risk assessor to rate endpoints in non-numerical terms (i.e., high, medium or low).

There is a tradeoff for the quantitative and qualitative approaches between flexibility and objectivity. The quantitative method requires the use of numerical scaling values to indicate the relative importance of each attribute. It is more systematic and requires substantially less case-by-case professional judgement and if the generic scaling values, such as those proposed in this paper, are applied. Assigning numerical scaling values to the endpoint attributes clearly documents the risk assessor's professional judgements and makes the decision process more transparent to risk managers and the general public. Determining a numerical score for each measurement endpoint using a previously established procedure may enable risk assessors and regulators to draw a

conclusion about risk in situations where the measurement results are contradictory and where the interested parties hold differing views on environmental assessment and protection.

The qualitative approach is somewhat more flexible, in that it is more amenable to determining the relative importance of the attributes on a case-specific basis. The risk assessor may opt either to assign weights on a case-by-case basis or to assume that each attribute is of equal importance. Assigning case-specific weights to the attributes enables the risk assessor to consider the nature of the measurement endpoints in question. However, if attributes are assigned weights on a case-specific basis, it is extremely important for the risk assessor to document the rationale for the relative weight given to each attribute. Thus, determining the weights for a qualitative approach may be simpler than for a quantitative evaluation, but documenting the rationale and the decision process requires a more extensive effort.

In order to use a weight-of-evidence evaluation to meet the requirements of a regulatory program and to provide a basis for a regulatory decision, the risk assessor needs the concurrence of the risk manager on the basic approach. In some cases, the regulator may consider a quantitative approach more useful; in others, a qualitative approach may be preferred. Whether a quantitative or qualitative approach is used, a systematic weight-of-evidence evaluation is likely to promote a broader and clearer understanding of the judgements incorporated in the ecological risk assessment.

6.0 SUMMARY

This paper outlines a weight-of-evidence approach for assessing ecological risks. The approach is conducted throughout the assessment; it is not carried out "after the fact." The workgroup has defined weight-of-evidence as the process by which measurement endpoint(s) are related to an assessment endpoint to evaluate if there is a significant risk of harm to the environment. The approach is planned and initiated at the problem formulation stage and results are integrated at the risk characterization stage.

The approach is organized around three components:

1. Weight assigned to each measurement endpoint;
2. Magnitude of response in the measurement endpoint; and
3. Concurrence among measurement endpoints.

A quantitative methodology was developed for each of these three components. The overall intent of the approach is to make transparent and more objective the various professional judgements made by ecological risk assessors as they evaluate information. The quantitative approach includes methods for: 1) weighing the individual measurement endpoints by evaluating how well they score against a set of ten attributes, 2) determining whether harm or lack of harm is indicated and the magnitude of response, and 3) graphically displaying the measurement endpoints in a matrix so that concurrence can be examined.

A simpler qualitative approach is also discussed. Risk assessors may choose between the quantitative and qualitative methods based on the needs of the assessment. In general, the quantitative approach is more objective and defensible and the qualitative approach is simpler, but requires greater documentation due to the added subjectivity.

The workgroup has applied the method to several case studies and found that it works reasonably well and that risk managers and risk assessors alike agree on the conclusions. One of the most valuable lessons of these exercises is that the application of the method provides a good basis for evaluating the selected measurement endpoints and for discussing the strengths and limitations of the assessment in an objective manner.

7.0 REFERENCES

Suter, G.W. 1993. Ecological Risk Assessment. Lewis Publishers. 538 p.

U.S. Environmental Protection Agency. 1992. Framework for Ecological Risk Assessment. EPA/630/R-92/001. Risk Assessment Forum , Washington, D.C.

U.S. Environmental Protection Agency. 1994. Ecological Risk Assessment Guidance for Superfund: Process for Designing and Conducting Ecological Risk Assessments. Review Draft. Edison, NJ.