EPA/630/R-98/004
January 1999

Report of the Workshop on Selecting Input Distributions For Probabilistic Assessments

U.S. Environmental Protection Agency
New York, NY
April 21-22, 1998

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

NOTICE

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

This report was prepared by Eastern Research Group, Inc. (ERG), an EPA contractor (Contract No. 68-D5-0028, Work Assignment No. 98-06) as a general record of discussions during the Workshop on Selecting Input Distributions for Probabilistic Assessments. As requested by EPA, this report captures the main points and highlights of discussions held during plenary sessions. The report is not a complete record of all details discussed nor does it embellish, interpret, or enlarge upon matters that were incomplete or unclear. Statements represent the individual views of each workshop participant; none of the statements represent analyses by or positions of the Risk Assessment Forum or the EPA.

CONTENTS

CONTENTS (Continued)

SECTION ONE

INTRODUCTION

1.1 BACKGROUND AND PURPOSE

The U.S. Environmental Protection Agency (EPA) has long emphasized the importance of adequately characterizing uncertainty and variability in its risk assessments, and it continuously studies various quantitative techniques for better characterizing uncertainty and variability. Historically, Agency risk assessments have been deterministic (i.e., based on a point estimate), and uncertainty analyses have been largely qualitative. In May 1997, the Agency issued a policy on the use of probabilistic techniques in characterizing uncertainty and variability. This policy recognizes that probabilistic analysis tools like Monte Carlo analysis are acceptable provided that risk assessors present adequate supporting data and credible assumptions. The policy also identifies several implementation activities that are designed to help Agency assessors review and prepare probabilistic assessments.

To this end, EPA's Risk Assessment Forum (RAF) is developing a framework for selecting input distributions for probabilistic assessment. This framework emphasizes parametric distributions, estimations of the parameters of candidate distributions, and evaluations of the candidate distributions' quality of fit. A technical panel, convened under the auspices of the RAF, began work on the framework in the summer of 1997. In September 1997, EPA sought input on the framework from 12 experts from outside the Agency. The group's recommendations included:

- Expanding the framework's discussion of exploratory data analysis and graphical methods for assess the quality of fit.
- Discussing distinctions between variability and uncertainty and their implications.
- Discussing empirical distributions and bootstrapping.
- Discussing correlation and its implications.
- Making the framework available to the risk assessment community as soon as possible.

In response to this input, EPA initiated a pilot program in which the Research Triangle Institute (RTI) applied the framework for fitting distributions to data from EPA's Exposure Factors Handbook (EFH) (US EPA, 1996a). RTI used three exposure factors—drinking water intake, inhalation rate, and residence time—as test cases. Issues highlighted as part of this effort fall into two broad categories: (1) issues associated with the *representativeness* of the data, and (2) issues associated with using the *Empirical Distribution Function (EDF)* (or resampling techniques) versus using a theoretical *Parametric Distribution Function (PDF).*

In April 1998, the RAF organized a 2-day workshop, "Selecting Input Distributions for Probabilistic Assessments," to solicit expert input on these and related issues. Specific workshop goals included:

- # Discussing issues associated with the selection of probability distributions.
- # Obtaining expert input on measurements, extrapolations, and adjustments.
- # Discussing qualitatively how to make quantitative adjustments.

EPA developed two issue papers to serve as a focal point for discussions: "Evaluating Representativeness of Exposure Factors Data" and "Empirical Distribution Functions and Non-parametric Simulation." These papers which were developed strictly to prompt discussions during the workshop are found in Appendix A. Discussions during the 2-day workshop focused on technical issues, not policy. The experts discussed issues that would apply to any exposure data.

This workshop report is intended to serve as an information piece for Agency assessors who prepare or review assessments based on the use of probabilistic techniques and who work with various exposure data. This report does not represent Agency guidance. It simply attempts to capture the technical rigor of the workshop discussions and will be used to support further development and application of probabilistic analysis techniques/approaches.

1.2 WORKSHOP ORGANIZATION

The workshop was held on April 21 and 22, 1998, at the EPA Region 2 offices in New York City. The 21 participants, experts in exposure and risk assessment, included biologists, chemists, engineers, mathematicians, physicists, statisticians, and toxicologists, and represented industry, academia, state agencies, EPA, and other federal agencies. A limited number of observers also attended the workshop. The experts and observers are listed in Appendix B.

The workshop agenda is in Appendix C. Mr. McCabe (EPA Region 2), Steven Knott of the RAF, and Dr. H. Christopher Frey, workshop facilitator, provided opening remarks. Before discussions began, Ms. Jacqueline Moya and Dr. Timothy Barry of EPA summarized the two issue papers.

During the 2-day workshop, the technical experts exchanged ideas in plenary and four small group breakout sessions. Discussions centered on the two issue papers distributed for review and comment before the workshop. Detailed discussions focused primarily on the questions in the charge (Appendix D). "Brainwriting" sessions were held within the smaller groups. Brainwriting, an interactive technique, enabled the experts to document their thoughts on a topic and build on each others' ideas. Each small group captured the essence of these sessions and presented the main ideas to the entire group during plenary sessions. A compilation of notes from the breakout sessions are included in Appendix E. Following expert input, observers were allowed to address the panel with questions or comments. In addition to providing input at the workshop, several experts provided pre- and postmeeting comments, which are in Appendices F and G, respectively.

Section Two of this report contains the chairperson's summary of the workshop. Section Three highlights workshop opening remarks. Section Four summarizes Agency presentations of the two issue papers. Sections Five and Six describe expert input on the two main topic areas—representativeness and EDF/PDF issues. Speakers' presentation materials (overheads and supporting papers) are included in Appendix H.

SECTION TWO

CHAIRPERSON'S SUMMARY
Prepared by: H. Christopher Frey, Ph.D.

The workshop was comprised of five major sessions, three of which were devoted to the issue of representativeness and two to issues regarding parametric versus empirical distributions and goodness-of-fit. Each session began with a trigger question. For the three sessions on representativeness, there was discussion in a plenary setting, as well as discussions within four breakout groups. For the two sessions regarding selection of parametric versus empirical distributions and the use of goodness-of-fit tests, the discussions were conducted in plenary sessions.

2.1 REPRESENTATIVENESS

The first session covered three main questions, based on the portion of the workshop charge (Appendix D) requesting feedback on the representativeness issue paper. After some general discussion, the following three trigger questions were formulated and posed to the group:

1. What information is required to fully specify a problem definition?

2. What constitutes (lack of) representativeness?

3. What considerations should be included in, added to, or excluded from the checklists given in the issue paper on representativeness (Appendix A)?

The group was then divided into four breakout groups, each of which addressed all three of these questions. Each group was asked to use an approach known as "brainwriting." Brainwriting is intended to be a silent activity in which each member of a group at any given time puts thoughts down on paper in response to a trigger question. After completing an idea, a group member exchanges papers with another group member. Typically, upon reading what others have written, new ideas are generated and written down. Thus, each person has a chance to read and respond to what others have written. The advantages of brainwriting are that all participants can generate ideas simultaneously, there is less of a problem with domination of the discussion by just a few people, and a written record is produced as part of the process. A disadvantage is that there is less "interaction" with the entire group. After the brainwriting activity was completed, a representative of each group reported the main ideas to the entire group.

The experts generally agreed that before addressing the issue of representativeness, it is necessary to have a clear problem definition. Therefore, there was considerable discussion of what factors must be considered to ensure a complete problem definition. The most general requirement for a good problem definition, to which the group gave general assent, is to specify the "who, what, when, where, why, and

how." The "who" addresses the population of interest. "Where" addresses the spatial characteristics of the assessment. "When" addresses the temporal characteristics of the assessment. "What" relates to the specific chemicals and health effects of concern. "Why" and "how" may help clarify the previous matters. For example, it is helpful to know that exposures occur because of a particular behavior (e.g., fish consumption) when attempting to define an exposed population and the spatial and temporal extent of the problem. Knowledge of "why" and "how" is also useful later for proposing mitigation or prevention strategies. The group in general agreed upon these principles for a problem definition, as well as the more specific suggestions detailed in Section 5.1.1 of this workshop report.

In regard to the second trigger question, the group generally agreed that "representativeness" is context-specific. Furthermore, there was a general trend toward finding other terminology instead of using the term "representativeness." In particular, many the group concurred that an objective in an assessment is to make sure that it is "useful and informative" or "adequate" for the purpose at hand. The adequacy of an assessment may be evaluated with respect to considerations such as "allowable error" as well as practical matters such as the ability to make measurements that are reasonably free of major errors or to reasonably interpret information from other sources that are used as an input to an assessment. Adequacy may be quantified, in principle, in terms of the precision and accuracy of model inputs and model outputs. There was some discussion of how the distinction between variability and uncertainty relates to assessment of adequacy. For example, one may wish to have accurate predictions of exposures for more than one percentile of the population, reflecting variability. For any given percentile of the population, however, there may be uncertainty in the predictions of exposures. Some individuals pointed out that, because often it is not possible to fully validate many exposure predictions or to obtain input information that is free of error or uncertainty, there is an inherently subjective element in assessing adequacy. The stringency of the requirement for adequacy will depend on the purpose of the assessment. It was noted, for example, that it may typically be easier to adequately define mean values of exposure than upper percentile values of exposure. Adequacy is also a function of the level of detail of an assessment; the requirements for adequacy of an initial, screening-level calculation will typically be less rigorous than those for a more detailed analysis.

Regarding the third trigger question, the group was generally complimentary of the proposed checklists in the representativeness issue paper (see Appendix A). The group, however, had many suggestions for improving the checklists. Some of the broader concerns were about how to make the checklists context-specific, because the degree of usefulness of information depends on both the quality of the information and the purpose of the assessment. Some of the specific suggestions included using flowcharts rather than lists; avoiding overlap among the flowcharts or lists; developing an interactive Web-based flowchart that would be flexible and context-specific; and clarifying terms used in the issue paper (e.g., "external" versus "internal" distinction). The experts also suggested that the checklists or flowcharts encourage additional data collection where appropriate and promote a "value of information" approach to help prioritize additional data collection. Further discussion of the group's comments is given in Section 5.1.3.

2.2 SENSITIVITY ANALYSIS

The second session was devoted to issues encapsulated in the following trigger questions:

How can one do sensitivity analysis to evaluate the implications of non-representativeness? In other words, how do we assess the importance of non-representativeness?

The experts were asked to consider data, models, and methods in answering these questions. Furthermore, the group was asked to keep in mind that the charge requested recommendations for immediate, short-term, and long-term studies or activities that could be done to provide methods or examples for answering these questions.

There were a variety of answers to these questions. A number of individuals shared the view that non-representativeness may not be important in many assessments. Specifically, they argued that many assessments and decisions consider a range of scenarios and populations. Furthermore, populations and exposure scenarios typically change over time, so that if one were to focus on making an assessment "representative" for one point in time or space, it could fail to be representative at other points in time or space or even for the original population of interest as individuals enter, leave, or change within the exposed population. Here again the notion of adequacy, rather than representativeness, was of concern to the group.

The group reiterated that representativeness is context-specific. Furthermore, there was some discussion of situations in which data are collected for "blue chip" distributions that are not specific to any particular decision. The experts did recommend that, in situations where there may be a lack of adequacy of model predictions based on available information, the sensitivity of decisions should be evaluated under a range of plausible adjustments to the input assumptions. It was suggested that there may be multiple tiers of analyses, each with a corresponding degree of effort and rigor regarding sensitivity analyses. In a "first-tier" analysis, the use of bounding estimates may be sufficient to establish sensitivity of model predictions with respect to one or more model outputs, without need for a probabilistic analysis. After a preliminary identification of sensitive model inputs, the next step would typically be to develop a probability distribution to represent a plausible range of outcomes for each of the sensitive inputs. Key questions to be considered are whether to attempt to make adjustments to improve the adequacy or representativeness of the assumptions and/or whether to collect additional data to improve the characterization of the input assumptions.

One potentially helpful criterion for deciding whether data are adequate is to try to answer the question: "Are the data good enough to replace an assumption?" If not, then additional data collection is likely to be needed. One would need to assess whether the needed data can be collected. A "value of information" approach can be useful in prioritizing data collection and in determining when sufficient data have been collected.

There was some discussion of sensitivity analysis of uncertainty versus sensitivity analysis of variability. The experts generally agreed that sensitivity analysis to identify key sources of uncertainty is a useful and appropriate thing to do. There was disagreement among the experts regarding the meaning of identifying key sources of variability. One expert argued that identifying key sources of variability is not useful, because variability is irreducible. However, knowledge of key sources of variability can be useful in identifying key characteristics of highly exposed subpopulations or in formulating prevention or mitigation measures. Currently, there are many methods that exist for doing sensitivity analysis, including running models for alternative scenarios and input assumptions and the use of regression or statistical methods to identify the most sensitive input distributions in a probabilistic analysis. In the short-term and long-term, it was suggested that some efforts be devoted to the development of "blue chip" distributions for

quantities that are widely used in many exposure assessments (e.g., intake rates of various foods). It was also suggested that new methods for sensitivity analysis might be obtained from other fields, with specific examples based on classification schemes, time series, and "g-estimation."

2.3 MAKING ADJUSTMENTS TO IMPROVE REPRESENTATION

In the third session, the group responded to the following trigger question:

How can one make adjustments from the sample to better represent the population of interest?

The group was asked to consider "population," spatial, and temporal characteristics when considering issues of representativeness and methods for making adjustments. The group was asked to provide input regarding exemplary methods and information sources that are available now to help in making such adjustments, as well as to consider short-term and long-term research needs.

The group clarified some of the terminology that was used in the issue paper and in the discussions. The term "population" was defined as referring to "an identifiable group of people." The experts noted that often one has a sample of data from a "surrogate population," which is not identical to the "target population" of interest in a particular exposure assessment. The experts also noted that there is a difference between the "analysis" of actual data pertaining to the target population and "extrapolation" of information from data for a surrogate population to make inferences regarding a target population. It was noted that extrapolation always "introduces" uncertainty.

On the temporal dimension, the experts noted that, when data are collected at one point in time and are used in an assessment aimed at a different point in time, a potential problem may occur because of shifts in the characteristics of populations between the two periods.

Reweighting of data was one approach that was mentioned in the plenary discussion. There was a discussion of "general" versus mechanistic approaches for making adjustments. The distinction here was that "general" approaches might be statistical, mathematical, or empirical in their foundations (e.g., regression analysis), whereas mechanistic approaches would rely on theory specific to a particular problem area (e.g., a physical, biological, or chemical model). It was noted that temporal and spatial issues are often problem-specific, which makes it difficult to recommend universal approaches for making adjustments. The group generally agreed that it is desirable to include or state the uncertainties associated with extrapolations. Several participants strongly expressed the view that "it is okay to state what you don't know," and there was no disagreement on this point.

The group recommended that the basis for making any adjustments to assumptions regarding populations should be predicated on stakeholder input and the examination of covariates. The group noted that methods for analyzing spatial and temporal aspects exist, if data exist. Of course, a common problem is scarcity of data and a subsequent reliance on surrogate information. For assessment of spatial variations, methods such as kreiging and random fields were commonly suggested. For assessment of temporal variations, time series methods were suggested.

There was a lively discussion regarding whether adjustments should be "conservative." Some experts initially argued that, to protect public health, any adjustments to input assumptions should tend to be biased in a conservative manner (so as not to make an error of understating a health risk, but with some nonzero probability of making an error of overstating a particular risk). After some additional discussion, it appeared that the experts were in agreement that one should strive primarily for accuracy and that ideally any adjustments that introduce "conservatism" should be left to decision makers. It was pointed out that invariably many judgments go into the development of input assumptions for an analysis and that these

judgments in reality often introduce some conservatism. Several pointed out that "conservatism" can entail significant costs if it results in over control or misidentification of important risks. Thus, conservatism in individual assessments may not be optimal or even conservative in a broader sense if some sources of risk are not addressed because others receive undue attention. Therefore, the overall recommendation of the experts regarding this issue is to strive for accuracy rather than conservatism, leaving the latter as an explicit policy issue for decision makers to introduce, although it is clear that individual participants had somewhat differing views.

The group's recommendations regarding measures that can be taken now include the use of stratification to try to reduce variability and correlation among inputs in an assessment, brainstorming to generate ideas regarding possible adjustments that might be made to input assumptions, and stakeholder input for much the same purpose, as well as to make sure that no significant pathways or scenarios have been overlooked. It was agreed that "plausible extrapolations" are reasonable when making adjustments to improve representativeness or adequacy. What is "plausible" will be context-specific.

In the short term, the experts recommended that the following activities be conducted:

Numerical Experiments. Numerical experiments can be used to test existing and new methods for making adjustments based on factors such as averaging times or averaging areas. For example, the precision and accuracy of the Duan-Wallace model (described in the representativeness issue paper in Appendix A) for making adjustments from one averaging time to another can be evaluated under a variety of conditions via numerical experiments.

Workshop on Adjustment Methods. The experts agreed in general that there are many potentially useful methods for analysis and adjustment but that many of these are to be found in fields outside the risk analysis community. Therefore, it would be useful to convene a panel of experts from other fields for the purpose of cross-disciplinary exchange of information regarding methods applicable to risk analysis problems. For example, it was suggested that geostatistical methods should be investigated.

Put Data on the Web. There was a fervent plea from at least one expert that data for "blue chip" and other commonly used distributions be placed on the Web to facilitate the dissemination and analysis of such data. A common concern is that often data are reported in summary form, which makes it difficult to analyze the data (e.g., to fit distributions). Thus, the recommendation includes the placement of actual data points, and not just summary data, on publicly accessible Web sites.

Suggestions on How to Choose a Method. The group felt that, because of the potentially large number of methods and the need for input from people in other fields, it was unrealistic to provide recommendations regarding specific methods for making adjustments. However, they did suggest that it would be possible to create a set of criteria regarding desirable features for such methods that could help an assessor when making choices among many options.

In the longer term, the experts recommend that efforts be directed at more data collection, such as improved national or regional surveys, to better capture variability as a function of different populations, locations, and averaging times. Along these lines, specific studies could be focused on the development or refinement of a select set of "blue chip" distributions, as well as targeted at updating or extending existing data sets to improve their flexibility for use in assessments of various populations, locations, and averaging

times. The group also noted that because populations, pathways, and scenarios change over time, there will be a continuing need to improve existing data sets.

2.4 EMPIRICAL AND PARAMETRIC DISTRIBUTION FUNCTIONS

In the fourth session, the experts began to address the second main set of issues as given in the charge. The trigger question used to start the discussion was:

> What are the primary considerations in choosing between the use of parametric distribution functions (PDFs) and empirical distribution functions (EDFs)?

The group was asked to consider the advantages of using one versus the other, whether the choice is merely a matter of preference, whether one is preferred, and whether there are cases when neither should be used.

The initial discussion involved clarification of the difference between the terms EDF and "bootstrap." Bootstrap simulation is a general technique for estimating confidence intervals and characterizing sampling distributions for statistics, as described by Efron and Tibshirani (1993). An EDF can be described as a stepwise cumulative distribution function or as a probability density function in which each data point is assigned an equal probability. Non-parametric bootstrap can be used to quantify sampling distributions or confidence intervals for statistics based upon the EDF, such as percentiles or moments. Parametric bootstrap methods can be used to quantify sampling distributions or confidence intervals for statistics based on PDFs. Bootstrap methods are also often referred to as "resampling" methods. However, "bootstrap" and EDF are not the same thing.

The experts generally agreed that the choice of EDF versus PDF is usually a matter of preference, and they also expressed the general opinion that there should be no rigid guidance requiring the use of one or the other in any particular situation. The group briefly addressed the notion of consistency. While consistency in the use of a particular method (e.g., EDF or PDF in this case) may offer benefits in terms of simplifying analyses and helping decision makers, there was a concern that any strict enforcement of consistency will inhibit the development of new methods or the acquisition of new data and may also lead to compromises from better approaches that are context-specific. Here again, it is important to point out that the experts explicitly chose not to recommend the use of either EDF or PDF as a single preferred approach but rather to recommend that this choice be left to the discretion of assessors on a case-by-case basis. For example, it could be reasonable for an assessor to include EDFs for some inputs and PDFs for others even within the same analysis.

Some participants gave examples of situations in which they might prefer to use an EDF, such as: (a) when there are a large number of data points (e.g., 12,000); (b) access to high speed data storage and retrieval systems; (c) when there is no theoretical basis for selecting a PDF; and/or (d) when one has an "ideal" sample. There was some discussion of preference for use of EDFs in "data-rich" situations rather than "data-poor" situations. However, it was noted that "data poor" is context-specific. For example, a data set may be adequate for estimating the 90th percentile but not the 99th percentile. Therefore, one may be "data rich" in the former case and "data poor" in the latter case with the same data set.

Some experts also gave examples of when they would prefer to use PDFs. A potential limitation of conventional EDFs is that they are restricted to the range of observed data. In contrast, PDFs typically provide estimates of "tails" of the distribution beyond the range of observed data, which may have intuitive or theoretical appeal. PDFs are also preferred by some because they provide a compact representation of data and can provide insight into generalizable features of a data set. Thus, in contrast to the proponent of the use of an EDF for a data set of 12,000, another expert suggested it would be easier to summarize the data with a PDF, as long as the fit was reasonable. At least one person suggested that a PDF may be easier to defend in a legal setting, although there was no consensus on this point.

For both EDFs and PDFs, the issue of extrapolation beyond the range of observed data received considerable discussion. One expert stated that, the "further we go out in the tails, the less we know," to which another responded, "when we go beyond the data, we know nothing." As a rebuttal, a third expert asked "do we really know nothing beyond the maximum data point?" and suggested that analogies with similar situations may provide a basis for judgments regarding extrapolation beyond the observed data. Overall, most or all of the experts appeared to support some approach to extrapolation beyond observed data, regardless of whether one prefers an EDF or a PDF. Some argued that one has more control over extrapolations with EDFs, because there are a variety of functional forms that can be appended to create a "tail" beyond the range of observed data. Examples of these are described in the issue paper. Others argued that when there is a theoretical basis for selecting a PDF, there is also some theoretical basis for extrapolating beyond the observed data. It was pointed out that one should not always focus on the "upper" tail; sometimes the lower tail of a model input may lead to extreme values of a model output (e.g., such as when an input appears in a denominator).

There was some discussion of situations in which neither an EDF or a PDF may be particularly desirable. One suggestion was that there may be situations in which explicit enumeration of all combinations of observed data values for all model inputs, as opposed to a probabilistic resampling scheme, may be desired. Such an approach can help, for example, in tracing combinations of input values that produce extreme values in model outputs. One expert suggested that neither EDFs nor PDFs are useful when there must be large extrapolations into the tails of the distributions.

A question that the group chose to address was, "How much information do we lose in the tails of a model output by not knowing the tails of the model inputs?" One comment was that it may not be necessary to accurately characterize the tails of all model inputs because the tails (or extreme values) of model outputs may depend on a variety of other combinations of model input values. Thus, it is possible that even if no effort is made to extrapolate beyond the range of observed data in model inputs, one may still predict extreme values in the model outputs. The use of scenario analysis was suggested as an alternative or supplement to probabilistic analysis in situations in which either a particular input cannot reasonably be assigned a probability distribution or when it may be difficult to estimate the tails of an important input distribution. In the latter case, alternative upper bounds on the distribution, or alternative assumptions regarding extrapolation to the tails, should be considered as scenarios.

Uncertainty in EDFs and PDFs was discussed. Techniques for estimating uncertainties in the statistics (e.g., percentiles) of various distributions, such as bootstrap simulation, are available. An example was presented for a data set of nine measurements, illustrating how the uncertainty in the fit of a parametric distribution was greatest at the tails. It was pointed out that when considering alternative PDFs (e.g., Lognormal vs. Gamma) the range of uncertainty in the upper percentiles of the alternative distributions will typically overlap; therefore, apparent differences in the fit of the tails may not be

particularly significant from a statistical perspective. Such insights are obtained from an explicit approach to distinguishing between variability and uncertainty in a "two-dimensional" probabilistic framework.

The group discussed whether mixture distributions are useful. Some experts were clearly proponents of using mixture distributions. A few individuals offered some cautions that it can be difficult to know when to properly employ mixtures. One example mentioned was for radon concentrations. One expert mentioned in passing that radon concentrations had been addressed in a particular assessment assuming a Lognormal distribution. Another responded that the concentration may more appropriately be described as a mixture of normal distributions. There was no firm consensus on whether it is better to use a mixture of distributions as opposed to a "generalized" distribution that can take on many arbitrary shapes. Those who expressed opinions tended to prefer the use of mixtures because they could offer more insight about processes that produced the data.

Truncation of the tails of a PDF was discussed. Most of the experts seemed to view this as a last resort fraught with imperfections. The need for truncation may be the result of an inappropriate selection of a PDF. For example, one participant asked, "If you truncate a Lognormal, does this invalidate your justification of the Lognormal?" It was suggested that alternative PDFs (perhaps ones that are less "tail heavy") be explored. Some suggested that truncation is often unnecessary. Depending upon the probability mass of the portion of the distribution that is considered for truncation, the probability of sampling an extreme value beyond a plausible upper bound may be so low that it does not occur in a typical Monte Carlo simulation of only a few thousand iterations. Even if an unrealistic value is sampled for one input, it may not produce an extreme value in the model output. If one does truncate a distribution, it can potentially affect the mean and other moments of the distribution. Thus, one expert summarized the issue of truncation as "nitpicking" that potentially can lead to more problems than it solves.

2.5 GOODNESS-OF-FIT

The fifth and final session of the workshop was devoted to the following trigger question:

> On what basis should it be decided whether a data set is adequately fitted by a parametric distribution?

The premise of this session was the assumption that a decision had already been made to use a PDF instead of an EDF. While not all participating experts were comfortable with this assumption, all agreed to base the subsequent discussion on it.

The group agreed unanimously that visualization of both the data and the fitted distribution is the most important approach for ascertaining the adequacy of fit. The group in general seemed to share a view that conventional Goodness-of-Fit (GoF) tests have significant shortcomings and that they should not be the only or perhaps even primary methods for determining the adequacy of fit.

One expert elaborated that any type of probability plot that allows one to transform data so that they can be compared to a straight line, representing a perfect fit, is extremely useful. The human eye is generally good at identifying discrepancies from the straight line perfect fit. Another pointed out that visualization and visual inspection is routinely used in the medical community for evaluation of information

such as x-rays and CAT scans; thus, there is a credible basis for reliance on visualization as a means for evaluating models and data.

One of the potential problems with GoF tests is that they may be sensitive to imperfections in the fit that are not of serious concern to an assessor or a decision maker. For example, if there are outliers at the low or middle portions of the distribution, a GoF test may suggest that a particular PDF should be rejected even though there is a good fit at the upper end of the distribution. In the absence of a visual inspection of the fit, the assessor may have no insight as to why a particular PDF was rejected by a GoF test.

The power of GoF tests was discussed. The group in general seemed comfortable with the notion of overriding the results of a GoF test if what appeared to be a good fit, via visual inspection, was rejected by the test, especially for large data sets or when the imperfections are in portions of the distribution that are not of major concern to the assessor or decision maker. Some experts shared stories of situations in which they found that a particular GoF test would reject a distribution due to only a few "strange" data points in what otherwise appears to be a plausible fit. It was noted that GoF tests become increasingly sensitive as the number of data points increases, so that even what appear to be small or negligible "blips" in a large data set are sufficient to lead to rejection of the fit. In contrast, for small data sets, GoF tests tend to be "weak" and may fail to reject a wide range of PDFs. One person expressed concern that any strict requirement for the use of GoF tests might reduce incentives for data collection, because it is relatively easy to avoid rejecting a PDF with few data.

The basis of GoF tests sparked some discussion. The "loss functions" assumed in many tests typically have to do with deviation of the fitted cumulative distribution function from the EDF for the data set. Other criteria are possible and, in principle, one could create any arbitrary GoF test. One expert asked whether minimization of the loss function used in any particular GoF test might be used as a basis for choosing parameter values when fitting a distribution to the data. There was no specific objection, but it was pointed out that a degree-of-freedom correction would be needed. Furthermore, other methods, such as maximum likelihood estimation (MLE), have a stronger theoretical basis as a method for parameter estimation.

The group discussed the role of the "significance level" and the "p-value" in GoF tests. One expert stressed that the significance level should be determined in advance of evaluating GoF and that it must be applied consistently in rejecting possible fits. Others, however, suggested that the appropriate significance level would depend upon risk management objectives. One expert suggested that it is useful to know the p-value of every fitted distribution so that one may have an indication of how good or weak the fit may have been according to the particular GoF test.

SECTION THREE

OPENING REMARKS

At the opening session of the workshop, representatives from EPA Region 2 and the RAF welcomed members of the expert panel and observers. Following EPA remarks, the workshop facilitator described the overall structure and objectives of the 2-day forum, which this section summarizes.

3.1 WELCOME AND REGIONAL PERSPECTIVE

Mr. William McCabe, Deputy Director, Program Support Branch, Emergency and Remedial Response Division, U.S. EPA Region 2

William McCabe welcomed the group to EPA Region 2 and thanked everyone for participating in the workshop. He noted that, in addition to this workshop, Region 2 also hosted the May 1996 Monte Carlo workshop, which ultimately led to the release of EPA's May 1997 policy document on probabilistic assessment. He commented on how this 2-day workshop was an important followup to the May 1996 event. Mr. McCabe stressed that continued discussions on viable approaches to probabilistic assessments are important because site-specific decisions rest on the merit of the risk assessment. He stated that this type of workshop is an excellent opportunity for attendees to discuss effective methods and expressed optimism that workshop discussions would provide additional insight and answers to probabilistic assessment issues. Resolution of key probabilistic assessment issues, he noted, will help the region members as they review risk assessments using probabilistic techniques. He mentioned, for example, the ongoing Hudson River PCB study for which deterministic and probabilistic assessments will be performed. In that case, as in others, Mr. McCabe said it will be critical for Agency reviewers to put the results into the proper context and to validate/critically review probabilistic techniques employed by the contractor(s) for the Potentially Responsible Parties.

3.2 OVERVIEW AND BACKGROUND

Mr. Steve Knott, U.S. EPA, Office of Research and Development, Risk Assessment Forum

On behalf of the RAF, Steve Knott thanked Region 2 for hosting the workshop. Mr. Knott briefly explained how the RAF originated in the early 1980s and comprises approximately 30 scientists from EPA program offices, laboratories, and regions. One primary RAF function is to bring experts together to carefully study and help foster cross-agency consensus on tough risk assessment issues.

Mr. Knott described the following activities related to probabilistic analysis in which the RAF has been involved:

- Formation of the 1983 ad hoc technical panel on Monte Carlo analysis.
- May 1996 workshop on Monte Carlo analysis (US EPA, 1996b).
- Development of the guiding principles for Monte Carlo analysis (US EPA, 1997a)

- EPA's general probabilistic analysis policy (US EPA, 1997b).

Mr. Knott reiterated the Agency's perspective on probabilistic techniques, stating that "the use of probabilistic techniques can be a viable statistical tool for analyzing variability and uncertainty in risk assessment" (US EPA, 1997b). Mr. Knott highlighted Condition 5 (on which this workshop was based) of the eight *conditions for acceptance* listed in EPA's policy:

> Information for each input and output distribution is to be provided in the report. This includes tabular and graphical representations of the distributions (e.g., probability density function and cumulative distribution function plots) that indicate the location of any point estimates of interest (e.g., mean, median, 95th percentile). *The selection of distributions is to be explained and justified.* For both the input and output distributions, variability and uncertainty are to be differentiated where possible (US EPA, 1997b).

Mr. Knott referred to the recent RTI report, "Development of Statistical Distributions for Exposure Factors" (1998), which presents a framework for fitting distributions and applies the framework to three case studies.

Mr. Knott explained that the Agency is seeking input from workshop participants primarily in the following areas:

- Methods for fitting distributions to less-than-perfect data (i.e., data that are not perfectly representative of the scenario(s) under study).

- Using the EDF (or resampling techniques) versus the PDF.

These issues were the focus of the workshop. Mr. Knott noted that the workshop will enable EPA to receive input from experts, build on existing guidance, and provide Agency assessors additional insight. EPA will use the information from this workshop in future activities, including (1) developing or revising guidelines and models, (2) updating the Exposure Factors Handbook, (3) supporting modeling efforts, and (4) applying probabilistic techniques to dose-response assessment.

3.3 WORKSHOP STRUCTURE AND OBJECTIVES
Dr. H. Christopher Frey, Workshop Chair

Dr. Frey, who served as workshop chair and facilitator, reiterated the purpose and goals of the workshop. As facilitator, Dr. Frey noted, he would attempt to foster discussions that would further illuminate and support probabilistic assessment activities. Dr. Frey stated that workshop discussions would center on the two issue papers mentioned previously. He explained that the RTI report was provided to experts for background purposes only. While the RTI report was not the review subject for this workshop, Dr. Frey commented that it may provide pertinent examples.

The group's charge, according to Dr. Frey, was to advise EPA and the profession on *representativeness* and *distribution function* issues. Because a slightly greater need exists for discussing representativeness issues and developing new techniques in this area, Dr. Frey explained that this topic

would receive the greatest attention during the 2-day workshop. He reemphasized that the workshop would focus on technical issues, not policy issues.

Dr. Frey concluded his introductory remarks by stating that the overall goal of the workshop was to provide a framework for addressing technical issues that may be applied widely to different future activities (e.g., development of exposure factor distributions).

Workshop Structure and Expert Charge

Dr. Frey explained that the workshop would be structured around technical questions related to the two issue papers. Appendix D presents the charge provided to experts before the workshop, including specific questions for consideration and comment. The workshop material, Dr. Frey noted, is inherently technical. He, therefore, encouraged the experts to use plain language where possible. He also noted that the workshop was not intended to be a short course or tutorial. In introducing the key topics for workshop discussions, Dr. Frey highlighted the following, which he perceived as the most challenging issues and questions based on experts' premeeting comments:

> ***Representativeness.*** How should assessors address representativeness? What deviation is acceptable (given uncertainty and variability in data quality, how close will we come to answering the question)? How do assessors work representativeness into their problem definition (e.g., What are we asking? What form will the answer take?)
>
> ***Sensitivity***. How important is the potential lack of representativeness? How do we evaluate this?
>
> ***Adjustment.*** Are there reasonable ways to adjust or extrapolate in cases where exposure data are not representative of the population of concern?
>
> ***EDF/PDF.*** How do assessors choose between EDFs and theoretical PDFs? On what basis do assessors decide whether a data set is adequately represented by a fitted analytic distribution?

Dr. Frey encouraged participants to remember the following general questions as they discussed specific technical questions during plenary sessions, small group discussions, and brainwriting sessions:

- What do we know today that we can apply to answer the questions or provide guidance?
- What short-term studies (e.g., numerical experiments) could answer the question or provide additional guidance?
- What long-term research (e.g., greater than 18 months) may be needed to answer the question or provide additional guidance?

According to Dr. Frey, the answers to these questions will help guide Agency activities related to probabilistic assessments.

Dr. Frey also encouraged the group to consider what, if anything, is not covered in the issue papers, but is related to the key topics. He noted some of the following examples, which were communicated in the experts' premeeting comments:

- Role of expert judgment and Bayesian methods, especially in making adjustments.
- Is model output considered representative if all the inputs to the model are considered representative? This issues relates, in part, to whether or not correlations or dependencies among the input are properly addressed.
- Role of representativeness in a default or generic assessment.
- Role of the measurement process.

Lastly, Dr. Frey explained that the activities related to the workshop are public information. The workshop was advertised in the Federal Register and observers were welcomed. Time was set aside on both days of the workshop for observer questions and comments.

SECTION FOUR

ISSUE PAPER PRESENTATIONS

Two issue papers were developed to present the expert panelists with pertinent issues and to initiate workshop discussions. Prior to the plenary and small group discussions, EPA provided an overview of each paper. This section provides a synopsis of each presentation. The two issue papers are presented in Appendix A. The overheads are in Appendix H.

4.1 ISSUE PAPER ON EVALUATING REPRESENTATIVENESS OF EXPOSURE FACTORS DATA

Jacqueline Moya, U.S. EPA, NCEA, Washington, DC

Ms. Moya opened her overview by noting that, while exposure distributions are available in the Exposure Factors Handbook, there is still a need to fit distributions for these data. Ms. Moya noted that a joint NCEA-RTI pilot project in September 1997 was established to do this. She then discussed the purpose of the issue paper and the main topics she planned to cover (i.e., framework for inferences, components of representativeness, the checklists, and methods for improving representativeness). The purpose of the issue paper, Ms. Moya reminded the group, was to introduce concepts and to prompt discussions on how to evaluate representativeness and what to do if a sample is *not* representative.

Ms. Moya presented a flow chart (see Figure 1 in the issue paper) of the data-collection process for a risk assessment. If data collection is not possible, she explained, surrogate data must be identified. The next step is to ask whether the surrogate data represent the site or chemical. Ms. Moya pointed to Checklist I (Assessing Internal Representativeness), which includes suggested questions for determining whether the surrogate data are representative of the population of concern. If not, the assessor must ask, "How do we adjust the data to make it more representative?"

Ms. Moya then briefly reviewed the key terms in the paper. *Representativeness* in the context of an exposure/risk assessment refers to the comfort with which one can draw inferences from the data. *Population* is defined in terms of its member characteristics (i.e., demographics, spatial and temporal elements, behavioral patterns). The assessor's *population of concern* is the population for which the assessment is being conducted. The *surrogate population* is the population used when data on the population of concern is not available. The *population of concern for the surrogate study* is the sample population for which the surrogate study was designed. The *population sampled* is a sample from the population of concern of the surrogate study.

Ms. Moya briefly described the external and internal components of representativeness. She explained that external components reflect how well the surrogate population represents the population of concern. Internal components refer to the surrogate study, specifically:

1. How well do sampled individuals represent the surrogate population? This depends on how well the study was designed. For example, was it random?

2. How well do the respondents represent the sample population? For example, if recreational fishermen are surveyed, is someone who fishes more frequently more likely to respond the survey, and therefore bias the response?

3. How well does the measured value represent the true value for the measurement unit? For example, are the recreational fishermen in the previous example accurately reporting the sizes of the fish they catch?

Ms. Moya reviewed the four checklists in the issue paper which may serve as tools for risk assessors trying to evaluate data representativeness. One checklist is for the population sampled versus the population of concern for the surrogate study (internal representativeness). The other checklists refer to the surrogate population versus the population of concern based on individual, spatial, and temporal characteristics (external representativeness). One goal of the workshop, Ms. Moya explained, was to solicit input from experts on the use of these checklists. Specifically, she asked whether certain questions should be eliminated (e.g., only a subset of the questions may be needed for a screening risk assessment).

Lastly, Ms. Moya pointed to discussions in the issue paper on attempting to improve representativeness. One section refers to how to make *adjustments* for differences in population characteristics (with discussions geared toward using weights for the sample). The second section refers to time-unit differences and includes how to adjust for this. Ms. Moya asked the group to consider how to evaluate the significance of population differences and how to perform extrapolations if they are necessary.

4.2 ISSUE PAPER ON EMPIRICAL DISTRIBUTION FUNCTIONS AND NON-PARAMETRIC SIMULATION

Timothy Barry, U.S. EPA, NCEA, Washington, DC

Dr. Barry reviewed the issues of concern related to selecting and evaluating distribution functions. He explained that, assuming data are representative, the risk assessor has two methods for representing an exposure factor in a probabilistic analysis: *parametric* (e.g., a Lognormal, Gamma, or Weibull distribution) and *non-parametric* (i.e., use the sample data to define an EDF).

To illustrate how the EDF is generated, Dr. Barry presented equations and histograms (see Appendix H). The basic EDF properties were defined as follows:

- Values between any two consecutive samples, x_k and x_{k+1}, cannot be simulated, nor can values smaller than the sample minimum, x_1, or larger than the sample maximum, x_n, be generated (i.e., $x>x_1$ and $x<x_n$).

- The mean of the EDF equals the sample mean. The variance of the EDF mean is always smaller than the variance of the sample mean; it equals $(n-1)/n$ times the variance of the sample mean.

- Expected values of simulated EDF percentiles are equal to the sample percentiles.

- If the underlying distribution is skewed to the right (as are many environmental quantities), the EDF tends to underestimate the true mean and variance.

In addition to the basic EDF, Dr. Barry explained, the following variations exist:

- *Linearized EDF*. In this case, a linearized cumulative distribution pattern results. The linearized EDF linearly extrapolates between two observations.

- *Extended EDF*. An extended EDF involves linearization and adds lower and upper tails to the data to reflect a "more realistic range" of the exposure variable. Tails are added based expert judgment.

- *Mixed Exponential*. In this case, an exponential upper tail is added to the EDF. This approach is based on extreme value theory.

After describing the basic concepts of EDFs, Dr. Barry provided an example in which investigators compared and contrasted parametric and non-parametric techniques. Specifically, 90 air exchange data points were shown to have a Weibull fit. When a basic EDF for these data is used, means and variance reproduce well. It was concluded that if the goal is to reproduce the sample, Weibull does well on the mean but poorly at the high end.

Dr. Barry encouraged the group to consider the following questions during the 2-day workshop:

- Is an EDF preferred over a PDF in any circumstances?

- Should an EDF not be used in certain situations?

- When an EDF is used, should the linearized, extended, or mixed version be used?

Dr. Barry briefly described the Goodness of Fit (GoF) questions the issue paper introduces. He explained that, generally, assessors should pick the simplest analytic distribution not rejected by the data. Because rejection depends on the chosen statistic and on an arbitrary level of statistical significance, Dr. Barry posed the following questions to the group:

- What role should the GoF statistic and its p-value (when available) play in deciding on the appropriate distribution?

- What role should graphical assessments of fit play?

- When none of the standard distributions fit well, should you investigate more flexible families of distributions (e.g., four parameter gamma, four parameter F, mixtures)?

SECTION FIVE

EVALUATING REPRESENTATIVENESS OF EXPOSURE FACTORS DATA

Discussions on the first day and a half of the workshop focused on developing a framework for characterizing and evaluating the *representativeness* of exposure data. The framework described in the issue paper on representativeness (see Appendix A) is organized into three broad sets of questions: (1) those related to differences in populations, (2) those related to differences in spatial coverage and scale, and (3) those related to differences in temporal scale. Therefore, discussions were held in the context of these three topic areas. The panel also discussed the strengths and weaknesses of the proposed "checklists" in the issue paper, which were designed to help the assessor evaluate representativeness. The last portion of the workshop session on representativeness included discussions on sensitivity (assessing the importance of non-representativeness) and on the methods available to adjust data to better represent the population of concern. This section describes the outcome of each of these discussions.

Initial deliberations centered on the need to define risk assessment objectives (i.e. problem definition) before evaluating the representativeness of exposure data. Discussions on sensitivity and adjustment followed.

5.1 PROBLEM DEFINITION

The group agreed on two points: that "representativeness" depends on the problem at hand and that the context of the risk analysis is critical. Several experts commented that assessors will have a difficult time defining representativeness if the problem has not been well-defined. The group therefore spent a significant amount of time discussing problem definition and problem formulation in the context of assessing representativeness. Several experts noted the importance of understanding the *end use* of the assessment (e.g., site-specific or generic, national or regional analysis). The group agreed that the most important step for assessors is to ask whether the data are representative enough for their intended use(s).

The group agreed that stakeholders and other data users should be involved in all phases of the assessment process, including early brainstorming sessions. Two experts noted that problem definition must address whether the assessment will adequately protect public health and the environment. Another expert stressed the importance of problem formulation, because not doing so risks running analyses or engaging resources needlessly. One participant commented that the importance of representativeness varies with the level (or tier) of the assessment. For example, if data are to be used in a screening manner, then conservativeness may be more important than representativeness. If data are to be used in something other than screening assessments, the assessor must consider *the value added* of more complex analyses (i.e., additional site-specific data collection, modeling). Two experts noted, however, that the following general problem statement/question would not change with a more or less sophisticated (tiered) assessment: Under an agreed upon set of exposure conditions, will the population of concern experience unacceptable risks? A more sophisticated analysis would merely enable a closer look at less conservative/more realistic conditions.

5.1.1 What information is required to specify a problem definition fully?

The group agreed that when defining any problem, the "fundamental who, what, when, where, why, and how" questions must be answered. One individual noted that if assessors answer these questions, they will be closer to determining if data are representative. The degree to which each basic question is important is specific to the problem or situation. Another reiterated the importance of remembering that the premier consideration is public health protection; he noted that if only narrow issues are discussed, the public health impact may be overlooked.

The group concurred that the problem must be defined in terms of location (space), time (over what duration and when in time), and population (person or unit). Some of these definitions may be concrete (e.g., spatial locations around a site), while some, like people who live on a brownfield site, may be more vague (e.g., because they may change with mobility and new land use). Because the problem addresses a future context, it must be linked to observable data by a model and assumptions. The problem definition should include these models and assumptions.

Various experts provided the following specific examples of the questions assessors should consider at the problem formulation stage of a risk assessment.

- What is the purpose of the assessment (e.g., regulatory decision, setting cleanup standards)?
- What is the population of interest?
- What type of assessment is being performed (site-specific or generic)?
- How is the assessment information being used? How will data be used (e.g., screening assessment versus court room)?
- Who are the stakeholders?
- What are the budget limitations? What is the cost/benefit of performing a probabilistic versus a deterministic assessment?
- What population is exposed, and what are its characteristics?
- How, when, and where are people exposed?
- In what activities does the exposed population engage? When does the exposed population engage in these activities, and for how long? Why are certain activities performed?
- What type of exposure is being evaluated (e.g., chronic/acute)?
- What is the scenario of interest (e.g., what is future land use)?
- What is the target or "acceptable" level of risk (e.g., 10^{-2} versus 10^{-6})?

- What is the measurement error?
- What is the acceptable level of error?
- What is the geographic scale and location (e.g., city, county)?
- What is the scale for data collection (e.g., regional/city, national)?
- What are site/region-specific issues (e.g., how might a warm climate or poor-tasting water affect drinking water consumption rates)?
- What is the temporal scale (day, year, lifetime)?
- What are the temporal characteristics of source emissions (continuous)?
- What is/are the route(s) of exposure?
- What is the dose (external, biological)?
- What is/are the statistic(s) of interest (e.g., mean, uncertainty percentile)?
- What is the plausible worst case?
- What is the overall data quality?
- What models must be used?
- What is the measurement error?
- When would results change a decision?

Many of the preceding questions are linked closely to defining representativeness. One subgroup compiled a list of key elements that are directly related to these types of questions when defining representativeness (see textbox on page 5-4).

5.1.2 What constitutes representativeness (or lack thereof)? What is "acceptable deviation"?

Several of the experts commented that, fundamentally, representativeness is a function of the quality of the data but reiterated that it depends ultimately on the overall assessment objective. Almost all data used in risk assessment fail to be representative in one or more ways. At issue is the effect of the lack of representativeness on the risk assessment. One expert suggested that applying the established concepts of EPA's data quality objective/data quality assessment process would help assessors evaluate data representativeness. Because populations are not fixed in time, one expert cautioned that if a data set is too representative, the risk assessment may be precise for only a moment. Another stressed the importance of taking a credible story to the risk manager. In that context, "precise representativeness" may be less

important than answering the question of whether we are being protective of public health. It is important

Sources of Variability and Uncertainty Related to the Assessment of Data Representativeness

EPA policy sets the standard that risk assessors should seek to characterize central tendency and plausible upper bounds on both individual risk and population risk for the overall target population as well as for sensitive subpopulations. To this extent, data representativeness cannot be separated from the assessment endpoint(s). Following are some key elements that may affect data representativeness. These elements are not mutually exclusive.

Exposed Population

- General target population
- Particular ethnic group
- Known sensitive subgroup (e.g., children, elderly, asthmatics)
- Occupational group (e.g., applicators)
- Age group (e.g., infant, child, teen, adult, whole life)
- Gender
- Activity group (e.g., sport fishermen, subsistence fishermen)

Geographic Scale, Location

- Trends (e.g., stationary, nonstationary behaviors)
- Past, present, future exposures
- Lifetime exposures
- Less-than-lifetime exposures (e.g., hourly, daily, weekly, annually)
- Temporal characteristics of source(s) (e.g., continuous, intermittent, periodic, concentrated, random)

Exposure Route

- Inhalation
- Ingestion (e.g., direct, indirect)
- Dermal (direct) contact (by activity; e.g., swimming)
- Multiple pathways

Exposure/Risk Assessment Endpoint

- Cancer risk
- Noncancer risk (margin of exposure, hazard index)
- Potential dose, applied dose, internal dose, biologically effective dose
- Risk statistic
- Mean, uncertainty percentile of mean
- Percentile of a distribution (e.g., 95th percentile risk)
- Uncertainty limit of variability percentile (upper confidence limit on 95th percentile risk)
- Plausible worst case, uncertainty percentile of plausible worst case

Data Quality Issues

- Direct measurement, indirect measurement (surrogates)
- Modeling uncertainties
- Measurement error (accuracy, precision, bias)
- Sampling error (sample size, non-randomness, independence)
- Monitoring issues (short-term, long-term, stationary, mobile)

to understand whether a lack of representativeness could mean the risk assessment results fail to protect public health or that they grossly overestimate risks.

One participant expressed concern that assessors feel deviations from representativeness can be measured. In reality, risk assessors may more often rely on qualitative or semiquantitative ways of describing that deviation. Another expert emphasized that assessors often have no basis on which to judge the representativeness of surrogate data (e.g., drinking water consumption), because rarely is local data available for comparison. Therefore, surrogate data, must be accepted or modified based on some qualitative information (e.g., the local area is hotter than that which the surrogate data is based).

The experts provided the following views on what constitutes representativeness and/or an acceptable level of non-representativeness. These views were communicated during small group and plenary discussions.

Nearly consistent with the definition in the issue paper, *representativeness* was defined by one subgroup as "the degree to which a value for a given endpoint *adequately* describes the value of that endpoint(s) likely seen in the target population." The term "adequately" replaces the terms "accurately and precisely" in the issue paper definition. One expert suggested changing the word representative to "useful and informative." The latter terms imply that one has learned something from the surrogate population. For example, the assessor may not prove the data are the same, but can, at minimum, capture the extent to which they differ. The term *non-representativeness* was defined as *"important differences* between target and surrogate populations with respect to the risk assessment objectives." Like others, this subgroup noted that the context of observation is important (e.g., what is being measured: environmental sample [water, air, soil] versus human recall [diet] versus tissue samples in humans [e.g., blood]). Assessors must ask about internal sample consistency, inappropriate methods, lack of descriptors (e.g., demographic, temporal), and inadequate sample size for targeted measure.

The group agreed, overall, that assessing adequacy or representativeness is inherently subjective. However, differing opinions were offered in terms of how to address this subjectivity. Several participants stressed the importance of removing subjectivity to the extent possible but without making future guidance too rigid. Others noted, however, that expert judgment is and must remain an integral part of the assessment process.

A common theme communicated by the experts was that representativeness depends on how much uncertainty and variability between the population of concern and the surrogate population the assessor is willing to accept. What is "good enough" is case specific, as is the "allowable error." Several experts commented that it is also important for assessors to know if they are comparing data means or tails. One expert suggested reviewing some case studies using assessments done for different purposes to illuminate the process of defining representativeness. "With regard to exposure factors, we [EPA] need to do a better job at specifying or providing better guidance on how to use the data that are available." For example, the soil ingestion data for children are limited, but they may suffice to provide an estimate of a mean. These data are not good enough to support a distribution or a good estimate of a high-end value, however.

One subgroup described representativeness/non-representativeness as the degree of bias between a data set and the problem. For example:

Scenario: Is a future residential scenario appropriate to the problem? For prospective risk assessment, there are usually irreducible uncertainties about making estimates about a future unknown population. Therefore, a certain amount of modeling must occur.

Model: Is a multiplicative, independent variable model appropriate? Uncertainties in the model can contribute to non-representativeness (e.g., it might not apply, it may be wrong, or calculations may be incorrect).

Variables: Is a particular study appropriate to the problem at hand—are the variables biased, uncertain? It may be easy to get confused about distinctions between bias (or inaccuracies), precision/imprecision, and representativeness/non representativeness. It is often assumed that a "representative" data set is one that has been obtained with a certain amount of randomization. More often, however, data that meet this definition are not available.

The group spokesperson explained that a well-designed and controlled randomized study yielding two results can be "representative" of the mean and dispersion but highly imprecise. Imprecision and representativeness are therefore different, but related. The central tendency of the distribution may be accurately estimated, but the upper percentile may not.

In summary, when assessing representativeness, the group agreed that emphasis should be placed on the *adequacy* of the data and how *useful and informative* a data set is to the defined problem. The group agreed that these terms are more appropriate than "accuracy and precision" in defining representative data in the context of a risk assessment. The importance of considering end use of the data was stressed and was a recurring theme in the discussions (i.e., how much representativeness is needed to answer the problem). Because the subject population is often a moving target with unpredictable direction in terms of its demographics and conditions of exposure, one expert commented that, in some cases, representativeness of a given data set may not be a relevant concept and generic models may be more appropriate.

5.1.3 What considerations should be included in, added to, or excluded from the checklists?

More than half the experts indicated that the checklists in Issue Paper 1 are useful for evaluating representativeness. One expert noted that regulators are often forced to make decisions without information. A checklist helps the assessor/risk manager evaluate the potential importance of missing exposure data. One expert re-emphasized the importance of allowing for professional judgement and expert elicitation when evaluating exposure data. Another panelist concurred, commenting that this type of the checklist is preferred over prescriptive guidance. Several of the experts noted, however, that checklists could be improved and offered several recommendations.

The group agreed that the checklist should be flexible for various problems and that users should be directed to consider the purpose of the risk assessment. The assessor must know the minimum requirements for a screening versus a probabilistic assessment. As one expert said, the requirements for a screening level assessment must differ from those for a full-blown risk assessment: Do I have enough information about the population (e.g., type, space, time) to answer the questions at this tier, and is that

information complete enough to make a management decision? Do I need to go through all the checklists before I can stop?

Instead of the binary (yes/no) and linear format of the checklists, several individuals suggested a flowchart format centered on the critical elements of representativeness (i.e., a "conditional" checklist)—to what extent does the representativeness of the data really matter? A flowchart would allow for a more iterative process and would help the assessor work through problem-definition issues. One expert suggested developing an interactive Web-based flowchart that would be flexible and context-specific. Another agreed, adding that criteria are needed to guide the assessor on what to do if information is not available. As one expert noted, questions should focus on the outcome of the risk assessment. The assessor needs to evaluate whether the outcome of the assessment changes if the populations differ.

One of the experts strongly encouraged collecting more/new data or information. Collection of additional data, he noted, is needed to improve the utility of these checklists. Another participant suggested that the user be alerted to the qualities of data that enable quantifying uncertainty and reminded that the degree of representativeness cannot be defined in certain cases. When biases due to lack of representativeness are suspected, how can assessors judge the direction of those biases?

In addition to general comments and recommendations, several individuals offered the following specific suggestions for the checklists:

- Clarifying definitions (e.g., internal versus external).

- Recategorizing. For example, use the following five categories: (1) interpreting measurements (more of a validity than representative issue), (2) evaluating whether sampling bias exists, (3) evaluating statistical sampling error, (4) evaluating whether the study measured what must be known, and (5) evaluating differences in the population. The first three issues are sources of internal error, the latter two are sources of external representativeness.

- Reducing the checklists. Several experts suggested combining Checklists II, III, and IV.

- Combining temporal, spatial, and individual categories. Avoid overlap in questions. For example, when overlap exists (e.g., in some spatial and temporal characteristics), which questions in the checklist are critical? A Web-based checklist, with the flow of questions appropriately programmed, could be designed to avoid duplication of questions.

- Including other populations of concern (e.g., ecological receptors).

- Including worked examples that demonstrate the criteria for determining if a question is answered adequately and appropriately. These examples should help focus the risk assessor on the issues that are critical to representativeness.

- Separating bias and sampling quality and extrapolation from reanalysis and reinterpretation.

- **#** Asking the following additional questions:

    - — Relative to application, is there consistency in the survey instruments used to collect the exposure data? How was measurement error addressed?

    - — Is the sample representative enough to bound the risk?

    - — Are data available on population characterization factors (e.g., age, sex)?

    - — What is known about the population of concern relative to the surrogate population? (If the population of concern is inadequately characterized, then the ability to consider the representativeness of the surrogate data is limited, and meaningless adjustment may result).

In summary, the group agreed on the utility of the checklists but emphasized the need to include in them decision criteria (i.e., how do we know if we have representative/non-representative data?) A brief discussion on the need to collect data followed. Some experts posed the following questions: How important is it to have more data? Is the risk assessment really driving decisions? Is more information needed to make good decisions? Is making risk assessment decisions on qualitative data acceptable? What data must to be collected, at minimum, to validate key assumptions? The results of the sensitivity analysis, as one expert pointed out, are key to answering these questions.

5.2 SENSITIVITY

How do we assess the importance of non-representativeness?

In considering the implications of non-representativeness, the group was asked to consider how one identifies the implications of non-representativeness in the context of the risk assessment. One expert commented that the term "non-representativeness" may be a little misleading, and as discussed earlier, finds the terms *data adequacy* or *data useability* more fitting to the discussions at hand. The expert noted that, from a Superfund perspective, data representativeness is only one consideration when assessing overall data quality or useability. Others agreed. The workshop chair encouraged everyone to discuss the suitability of the term "representativeness" while assessing its importance during the small group discussions.

One group described a way in which to assess the issue of non-representativeness as follows: The assessor must check the sensitivity of decisions to be made as a result of the assessment. That is, under a range of plausible adjustments, will the risk decision change? Representativeness is often not that important because risk management decisions depend on a range of target populations under various scenarios. A few of the experts expressed concern that problems will likely arise if the exposure assessor is separated from decision makers. One person noted that often times an exposure assessment will be done absent of a specific decision (e.g., nonsite, non-Superfund situation). Another noted that in the pesticide program situations occur in which an exposure assessment is done before toxicity data are available. Such separations may be unavoidable. Another expert emphasized that any future guidance should stress the importance of assessors being cognizant of data distribution needs even if the assessors are removed from the decision or have limited data.

One individual noted that examples would help. The assessor should perform context-specific sensitivity analysis. It would help to develop case studies and see how sensitivity analysis affects application (e.g., decision focus).

Another group discussed sensitivity analysis in the context of a tiered approach. For the first tier, a value that is "biased high" should be selected (e.g., 95th percentile upper bound). The importance of a parameter (as evidenced by a sensitivity analysis) is determined first, making the representativeness or non-representativeness of the nonsensitive parameters unimportant. For the second tier (for sensitive parameters), the assessor must consider whether averages or high end estimates are of greater importance. This group presented an example using a corn oil scenario to illustrate when differences between individuals (e.g. high end) and mixtures (averages) may be important. Because corn oil is a blend with input from many ears of corn, if variability exists in the contaminant concentrations in individual ears of corn, then corn oil will typically represent some type of average of those concentrations. For such a mixture, representativeness is less of an issue. It is not necessary to worry about peak concentrations in one ear of corn. Instead, one would be interested in situations which might give rise to a relatively high average among the many ears of corn that comprise a given quantity of corn oil. If one is considering individual ears of corn, it becomes more important to have a representative sample; the tail of the distribution becomes of greater interest.

A third subgroup noted that, given a model and parameters, assessors must determine whether enough data exist to bound the estimates. If they can bound the estimates, a sensitivity analysis is performed with the following considerations: (1) identify the sensitive parameters in the model; (2) focus on sensitive parameters and evaluate the distribution beyond the bounding estimate (i.e., identify the variability of these parameters) for the identified sensitive parameters; (3) evaluate whether the distribution is representative; and (4) evaluate whether more data should be collected or if an adjustment is appropriate.

Members of the remaining subgroup noted, and others agreed, that a "perfect" risk assessment is not possible. They reiterated that it is key to evaluate the data in the context of the decision analysis. Again, what are the consequences of being wrong, and what difference do decision errors make in the estimate of the parameter being evaluated? This group emphasized that the question is situation-specific. In addition, they noted the need for placing bounds on data used.

One question asked throughout these discussions was "Are the data good enough to replace an existing assumption and, if not, can we obtain such data?" One individual again stressed the need for "blue chip" distributions at the national level (e.g., inhalation rate, drinking water). Another expert suggested adding activity patterns to the list of needed data.

In summary, the group generally agreed that the sensitivity of the risk assessment decision must be considered before non-representativeness is considered problematic. In some cases, there may not be an immediate decision, but good distributions are still important.

How can one do sensitivity analysis to evaluate the implications of non-representativeness?

The workshop chair asked the group to consider the mechanics of a sensitivity analysis. For example, is there a specific statistic that should be used, or is it decision dependent? One expert responded by noting that sensitivity analysis can be equated to partial correlation coefficients (which are internal to a

model). He noted, however, that sensitivity analysis in the context of exposure assessment is more "bottom line" sensitivity (i.e., if an assumption is changed, how does the change affect the bottom line?). The focus here is more external—what happens when you change the inputs to the model (e.g., the distributions)? Another pointed to ways in which to perform internal sensitivity analysis. For example, the sensitivity of uncertainty can be separated out from the sensitivity of the variability component (see William Huber's premeeting comments on sensitivity). Another expert stressed, however, that sensitivity analysis is inherently tied to uncertainty; it is not tied to variability unless the variability is uncertain. It was noted that sensitivity analysis is an opportunity to view things that are subjective. Variability, in contrast is inherent in the data, unless there are too few data to estimate variability sufficiently. One expert commented that it is useful to know which sources of variability are most important in determining exposure and risk.

One individual voiced concern regarding how available models address sensitivity. Another questioned whether current software (e.g., Crystal Ball® and @Risk®) covers sensitivity coefficients adequately (i.e., does it reflect the depth and breadth of existing literature?).

Lastly, the group discussed sensitivity analysis in the context of what we know now and what we need to know to improve the existing methodology. Individuals suggested the following:

- Add the ability to classify sample runs to available software. Classify inputs and evaluate the effect on outputs.

- Crystal Ball® and @Risk® are reliable for many calculations, but one expert noted they may not currently be useful for second-order estimates, nor can they use time runs. Time series analyses are particularly important for Food Quality Protection Act (FQPA) evaluations.

- Consider possible biases built into the model due to residuals lost during regression analyses. This factor is important to the sensitivity of the model prediction.

One expert pointed out that regression analyses can introduce bias because residuals are often dropped out. Others agreed that this is an important issue. For example, it can make an order-of-magnitude difference in body weight and surface area scaling. Another expert stated that this issue is of special interest for work under the FQPA, where use of surrogate data and regression analysis is receiving more and more attention. Another expert noted that "g-estimation" looks at this issue. The group revisited this issue during their discussions on adjustment.

5.3 ADJUSTMENT

How can one adjust the sample to better represent the population of interest?

The experts addressed adjustment in terms of population, spatial, and temporal characteristics. The group was asked to identify currently available methods and information sources that enable the quantitative adjustment of surrogate sample data. In addition, the group was asked to identify both short- and long-term research needs in this area. The workshop chair noted that the issue paper only includes discussion on adjustments to account for time-scale differences. The goal, therefore, was to generate some

discussion on spatial and population adjustments as well. Various approaches for making adjustments were discussed, including general and mechanistic. General approaches include those that are statistically-, mathematically-, or empirically-based (e.g., regression analysis). Mechanistic approaches would involve applying a theory specific to a problem area (e.g., a biological, chemical, or physical model).

Some differing opinions were provided as to how reliably we can apply available statistics to adjust data. In time-space modeling, where primary data and multiple observations occur at different spatial locations or in multiple measures over time, one expert noted that a fairly well-developed set of analytic methods exist. These methods would fall under the category of mix models, kreiging studies for spatial analysis, or random-effects models. The group agreed that extrapolating or postulating models are less well-developed. One person noted that classical statistics fall short because they do not apply to situations in which representativeness is a core concern. Instead, these methods focus more on the accuracy or applicability of the model. The group agreed that statistical literature in this area is growing.

Another individual expressed concern that statistical tools and extrapolations introduce more uncertainty to the assessment. This uncertainty may not be a problem if the assessor has good information about the population of concern and is simply adjusting or reweighing the data, but when the assessor is extrapolating the source term, demographics, and spatial characteristics simultaneously, more assumptions and increasing uncertainty are introduced.

In general, the group agreed that a model-based approach has merit in certain cases. The modeled approach, as one expert noted, is a cheap and effective approach and likely to support informed/more objective decisions. The group agreed that validated models (e.g., spatial/fate and transport models) should be used. Because information on populations may simply be unavailable to validate some potentially useful models, several participants reemphasized the need to collect more data, which was a recurring workshop theme.

One expert pointed out that the assessor must ask which unit of observation is of concern. For example, when evaluating cancer risk, temporal/spatial issues (e.g., residence time) are less important. When evaluating developmental effects (when windows of time are important), however, the temporal/spatial issues are more relevant. Again, assessors must consider the problem at hand before identifying the unit of time.

From a pesticide perspective, it was noted that new data cannot always be required of registrants. When considering the effects of pesticides, for example, crop treatment rates change over time. As a result, bridging studies are used to link available application data to crop residues (using a multiple linear regression model).

One expert stressed the importance and need for assessors to *recognize uncertainty*. Practitioners of probabilistic assessment should be encouraged to aggressively evaluate and discuss the uncertainties in extrapolations and their consequences. Often, probabilistic techniques can provide better information for better management decisions. The expert pointed out that, in some cases, one may not be able to assign a distribution, or one may choose not to do so because it would risk losing valuable information. In those cases, multiple scenarios and results reported in a nonprobabilistic way (both for communication and management decisions) may be appropriate.

At this point, one expert suggested that the discussion of multiple scenarios was straying from the basic question to be answered— "If I have a data set that does not apply to my population, what do I need to do, if anything?" Others disagreed, noting that it may make sense to run different scenarios and evaluate the difference. If a different scenario makes a difference, more data must be collected. One expert argued, however, that we cannot wait to observe trends; assessors must predict the future based on a "snapshot" of today.

One expert suggested the following hierarchy when deciding on the need to refine/adjust data:

- Can the effect be bounded? If yes, no adjustment is needed.
- If the bias is conservative, no adjustment is needed.
- Use a simple model to adjust the data.
- If adjustments fail, resample/collect more data, if possible.

The group then discussed the approaches and methods that are currently available to address non-representative data, and indicated that the following approaches are viable:

1. Start with brainstorming. Obtain stakeholder input to determine how the target population differs from the population for which you have data.

2 Look at covariates to get an idea of what adjustment might be needed. Stratify data to see if correlation exists. Stratification is a good basis for adjustments.

3. Use "kreiging" techniques (deriving information from one sample to a smaller, sparser data set). Kreiging may not fully apply to spatial, temporal, and population adjustments, however, because it applies to the theory of random fields. Kreiging may help improve the accuracy of existing data, but it does not enable extrapolation.

4. Include time-steps in models to evaluate temporal trends.

5. Use the "plausible extrapolation" model. This model is acceptable if biased conservatively.

6. Consider spatial estimates of covariate data (random fields).

7 Use the scenario approach instead of a probabilistic approach.

.

8. Bayesian statistical methods may be applicable and relevant.

   One expert presented a brief case study as an example of Bayesian analysis of variability and uncertainty and use of a covariate probability distribution model based on regression to allow extrapolation to different target populations. The paper he summarized, "Bayesian Analysis of Variability and Uncertainty on Arsenic Concentrations in U.S. Public Water Supplies," and supporting overheads, are in Appendix G. The paper describes a Bayesian methodology for estimating the distribution and its dependence on

covariates. Posterior distributions were computed using Markov Chain Monte Carlo (MCMC). In this example, uncertainties and variability were associated with time issues and the self-selected nature of arsenic samples. After briefly reviewing model specifications and distributional assumptions, the results and interpretations were presented, including a presentation of MCMC output plots and the posterior cumulative distribution of source water. The uncertainty of fitting site-specific data to the national distribution of arsenic concentrations was then discussed. The results suggest that Bayesian methodology powerfully characterizes variability and uncertainty in exposure factors. The probability distribution model with covariates provides insights and a basis for extrapolation to other targeted populations or subpopulations. One of the main points of presenting this methodology was to demonstrate the use of covariates. This case study showed that you can fit a model with covariates, explicitly account for residuals (which may be important), and apply that same model to a separate subpopulation where you know something about the covariates. According to the presenter, such an approach helps reveal whether national data represent local data.

When evaluating research needs, one expert pointed out that assessors should identify the minimal amount of information they need to analyze the data using available tools. The group offered the following suggestions for both short and long-term research areas. The discussion of short-term needs also included recommendations for actions the assessors can take now or in the short term to address the topics discussed in this workshop.

Short-term research areas and actions

1. Design studies for data collection that are amenable to available methods for data analysis. Some existing methods are unusable because not all available data, which were used to support the methods, are from well-designed studies.

2. Validate existing models on population variability (e.g., the Duan-Wallace model [Wallace et al., 1994] and models described by Buck et al. [1995]). This validation can be achieved by collecting additional data.

3. Run numerical experiments to test existing and new methods for making adjustment based on factors such as averaging times or area. Explore and evaluate the Duan-Wallace model.

4. Hold a separate workshop on adjustment methods (e.g., geostatistical and time series methods). Involve the modelers working with these techniques on a cross-disciplinary panel to learn how particular techniques might apply to adjustment issues that are specific to risk assessment.

5. Provide guidelines on how to evaluate or choose an available method, instead of simply describing available techniques. These guidelines would help the assessor determine whether a method applies to a specific problem.

6. To facilitate their access and utility, place national data on the Web (e.g., 3-day CSFII data, 1994–1996 USDA food consumption data). Ideally, the complete data set, not just

summary data, could be placed on the Web because data in summary form is difficult to analyze (e.g., to fit distributions).

Possible long-term research areas

1. Collect additional exposure parameter data on the national and regional levels (e.g., "blue chip" distributions). One expert cautioned that some sampling data have been or may be collected by field investigators working independently of risk assessment efforts. Therefore, risk assessors should have input in methods for designing data collection.

2. Perform targeted studies (spatial/temporal characteristics) to update existing data.

Discussions of *adjustment* ended with emphasis on the fact that adjustment and the previously described methods *only need be considered if they impact the endpoint*. One expert reiterated that when no quantitative or objective ways exist to adjust the surrogate data, a more generalized screening approach should be used.

As a follow-up to the adjustment discussions, a few individuals briefly discussed the issue of "bias/loss function" to society. Because this issue is largely a policy issue, it only received brief attention. One expert noted that overconservatism is undesirable. Another stressed that it is not in the public interest to extrapolate in the direction of not protecting public health; assessors should apply conservative bias but make risk managers aware of the biases. The other expert countered that blindly applying conservative assumptions could result in suboptimal decisions, which should not be taken lightly. In general, the group agreed on the following point: Assessors should use their *best scientific judgment* and strive for accuracy when considering representativeness and uncertainty issues. Which choice will ensure protection of public health without unreasonable loss? It was noted that the cost of overconservatism should drive the data-collection push (e.g., encourage industry to contribute to data collection efforts because they ultimately pay for conservative risk assessments).

5.4 SUMMARY OF EXPERT INPUT ON EVALUATING REPRESENTATIVENESS

Workshop discussions on representativeness revealed some common themes. The group generally agreed that representativeness is context-specific. Methods must be developed to ensure representativeness exists in cases where lack of representativeness would substantially impact a risk-management decision. Methods, the sensitivity analysis, and the decision endpoint are closely linked. One expert warned that once the problem is defined, the assessor must understand how to use statistical tools properly to meet assessment goals. Blind application of these tools can result in wrong answers (e.g., examining the tail versus the entire curve).

One or more experts raised the following issues related to evaluating the quality and "representativeness" of exposure factors data:

- Representativeness might be better termed "adequacy" or "usefulness."

- Before evaluating representativeness, the risk assessor, with input from stakeholders, must define the assessment problem clearly.

- # No data are perfect; assessors must recognize this fact, clearly present it in their assessments, and adjust non-representative data as necessary using available tools. The assessors must make plausible adjustments if non-representativeness matters to the endpoint.

- # To perform a probabilistic assessment well, adequate data are necessary, even for an assessment with a well-defined objective. In large part, current exposure distribution data fall short of the risk assessors' needs. Barriers to collecting new data must be identified, then removed. Cost limitations were pointed out, however. One expert, therefore, recommended that justification and priorities be established.

- # Methods must be sensitive to needs broader than the Superfund/RCRA programs (e.g., food quality and pesticide programs).

- # When evaluating the importance of representativeness and/or adjusting for non-representativeness, the assessor needs to make decisions that are adequately protective of public health while still considering costs and other loss functions. Ultimately, the assessor should strive for accuracy.

Options for the assessor when the population of concern has been shown to have different habits than the surrogate population were summarized as follows: (1) determine that the data are clearly not representative and cannot be used; (2) use the surrogate data and clearly state the uncertainties; or (3) adjust the data, using what information is available to enable a reasonable adjustment.

SECTION SIX

EMPIRICAL DISTRIBUTION FUNCTIONS AND RESAMPLING VERSUS PARAMETRIC DISTRIBUTIONS

Assessors often must understand and judge the use of parametric methods (e.g., using such theoretical distribution functions as the Lognormal, Gamma, or Weibull distribution) versus non-parametric methods (using an EDF) for a given assessment. The final session of the workshop was therefore dedicated to exploring the strengths and weaknesses of EDFs and issues related to judging the quality of fit for theoretical distributions. Discussions centered largely on the topics in Issue Paper 2 (see Appendix A for a copy of the paper and Section 3 for the workshop presentation of the paper). This section presents a summary of expert input on these topics.

Some of the experts thought the issue paper imposed certain constraints on discussions because it assumed that: (1) no theoretical premise exists for assuming a parametric distribution, and (2) the data are representative of the exposure factor in question (i.e., obtained as a simple random sample and in the proper scale). These experts noted that many of the assertions in the issue paper do not exist in reality. For example, it is unlikely to find a perfectly random sample for exposure parameter data.

As a result, the discussions that followed covered the relative advantages and disadvantages of parametric and non-parametric distributions under a broader range of conditions.

6.1 SELECTING AN EDF OR PDF

Experts were asked to consider the following questions.

What are the primary considerations in choosing between the use of EDFs and theoretical PDFs? What are the advantages of one versus the other? Is the choice a matter of preference? Are there situations in which one method is preferred over the other? Are there cases in which neither method should be used?

The group agreed that selecting an EDF versus a PDF is often a matter of personal preference or professional judgment. It is not a matter of systematically selecting either a PDF- or EDF-based approach for every input. It was emphasized that selection of a distribution type is case- or situation-specific. In some cases, both approaches might be used in a single assessment. The decision, as one expert pointed out, is driven largely by data-rich versus data-poor situations. The decision is based also on the risk assessment objective. Several experts noted that the EDF and PDF have different strengths in different situations and encouraged the Agency not to recommend the use of one over the other or to develop guidance that is too rigid. Some experts disputed the extent to which a consistent approach should be encouraged. While it was recognized that a consistent approach may benefit decision making, the overall consensus was that too many constraints would inhibit the implementation of new/innovative approaches, from which we could learn.

Technical discussions started with the group distinguishing between "bootstrap" methods and EDFs. One expert questioned if the methods were synonymous. EDF, as one expert explained, is a

specific type of step-wise distribution that can be used as a basis for bootstrap simulations. EDF is one way to describe a distribution using data; bootstrapping enables assessors to resample that distribution in a special way (e.g., setting boundaries on the distribution of the mean or percentile) (Efron and Tibshirani, 1993). Another expert distinguished between a parametric and non-parametric bootstrap, stating that there are good reasons for using both methods. These reasons are well-covered in the statistical literature. One expert noted that bootstrapping enables a better evaluation of the uncertainty of the distribution.

Subsequent discussion focused on expert input on deciding which distribution to fit, if any, for a given risk assessment problem. That is, if the assessor has a data set that must be represented, is it better to use the data set as is and not make any assumptions or to fit the data set to a parametric distribution? The following is a compilation of expert input.

- *Use of the EDF*. The use of an EDF may be preferable (1) when a large number of data points exists, (2) when access is available to computers with high speed and storage capabilities, (3) when no theoretical basis for selecting a PDF exists, or (4) when a "perfect" data set is available. With small data sets, it was noted that the EDF is unlikely to represent an upper percentile adequately; EDFs are restricted to the range of observed data. One expert stated that while choice of distribution largely depends on sample size, in most cases he would prefer the EDF.

  When measurement or response error exists, one expert pointed out that an EDF should not be used before looking at other options.

- *Use of the PDF*. One expert noted that it is easier to summarize a large data set with a PDF as long as the fit is reasonable. Use of PDFs can provide estimates of "tails" of the distribution beyond the range of observed data. A parametric distribution is a convenient way to concisely summarize a data set. That is, instead of reporting the individual data values, one can report the distribution and estimated parameter values of the distribution.

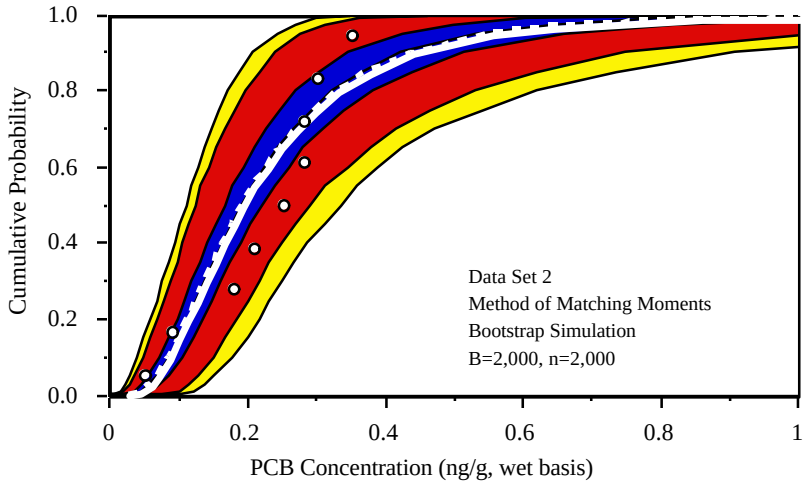
  While data may not be generated exactly according to a parametric distribution, evaluating parametric distributions may provide insight to generalizable features of a data set, such as moments, parameter values, or other statistics. Before deciding which distribution to use, two experts pointed out the value of trying to fit a parametric distribution to gain some insight about the data set (e.g., how particular parameters may be related to other aspects of the data set). These experts felt there is great value in examining larger data sets and thinking about what tools can be used to put data into better perspective. Another expert noted that the PDF is easier to defend at a public meeting or in a legal setting because it has some theoretical basis.

- *Assessing risk assessment outcome*. The importance of understanding what the implications of the distribution choice are to the outcome of the risk assessment was stressed. An example of fitting soil ingestion data to a number of parametric and non-parametric distributions yielded very different results. Depending on which distribution was used, cleanup goals were changed by approximately 2 to 3 times. Therefore, the choice may have cost implications.

Assuming all data are empirical. One expert felt strongly that all distributions are empirical. In data poor situations, why assume that the data are Lognormal? The data could be bimodal in the tails. If a data set is assumed to be empirical, there is some control as to how to study the tails. Another expert reiterated that using EDFs in data poor situations (e.g., six data points) does not enable simulation above or below known data values. One expert disagreed providing an example that legitimizes the concern for assuming that data fit a parametric distribution. He noted that if there is no mechanistic basis for fitting a parametric distribution, and a small set of data points by chance are at the lower end of the distribution, the 90th percentile estimate will be wrong.

Evaluating uncertainty. Techniques for estimating uncertainty in EDFs and PDFs were discussed. The workshop chair presented an example in which he fit a distribution for variability to nine data points. He then placed uncertainty bands around the distributions (both Normal and Lognormal curves) using parametric bootstrap simulation. (See Figure 6-1). For example, bands were produced by plotting the results of 2,000 runs of a synthetic data set of nine points sampled randomly from the Lognormal distribution fitted to the original data set. The wide uncertainty (probability) bands indicate the confidence in the distribution. This is one approach for quantifying how much is known about what is going on at the tails, based on random sampling error. When this exercise was performed for the Normal distribution, less uncertainty was predicted in the upper tail; however, a lower tail with negative values was predicted, which is not appropriate for a non-negative physical quantity such as concentrations. The chair noted that, if a stepwise EDF had been used, high and low ends would be truncated and tail concentrations would not have been predicted. This illustrates that the estimate of uncertainty in the tails depends on which assumption is made for the underlying distribution. Considering uncertainty in this manner allows the assessor to evaluate alternative distributions and gain insight on distinguishing between variability and uncertainty in a "2-dimensional probabilistic framework." Several participants noted that this was a valuable example.

Figure 6-1: Variability and Uncertainty in the Fit of Lognormal Distribution to a Data Set of n=9 (Frey, H.C. and D.E. Burmaster, 1998)



Extrapolating beyond the range of observable data. The purpose of the risk analysis drives what assessors must know about the tails of the distribution. One expert emphasized that the further assessors go into the tails, the less they know. Another stressed that once assessors get outside the range of the data, they know nothing. Another expert disagreed with the point that assessors know nothing beyond the highest data point. He suggested using analogous data sets that are more data rich to help in predicting the tails of the distribution. The primary issue becomes how much the assessors are willing to extrapolate.

Several experts agreed that uncertainty in the tails is not always problematic. If the assessor wants to focus on a subgroup, for example, it is not necessary to look at the tail of the larger group. *Stratification*, used routinely by epidemiologist, was suggested. With stratification, the assessor would look at the subgroup and avoid having to perform an exhaustive assessment of the tail, especially for more preliminary calculations used in a tiered approach. In a tiered risk assessment system, if the assessor assumes the data are Lognormal, standard multiplicative equations can be run on a simple calculator. While Monte Carlo-type analyses can provide valuable information in many cases, several experts agreed that probabilistic analyses are not always appropriate or necessary. It was suggested that, in some cases, deterministic scenario-based analyses, rather than Monte Carlo simulation, would be a useful way to evaluate extreme values for a model output.

In a situation where a model is used to make predictions of some distribution, several experts agreed that the absence of perfect information about the tails of the distribution of each input does not mean that assessors will not have adequate information about the tail of the model output. Even if all we have is good information about the central portions of the input distributions, it may be possible to simulate an extreme value for the model output.

Use of data in the tails of the distribution. One expert cautioned assessors to be sensitive to potentially important data in the tails. He provided an example in which assessors relied on the "expert judgement" of facility operators in predicting contaminant releases from a source. They failed to adequately predict "blips" that were later shown to exist in 20 to 30 percent of the distribution. Another expert noted that he was skeptical about adding tails (but was not skeptical about setting upper and lower bounds). It was agreed that, in general, assessors need to carefully consider what they do know about a given data set that could enable them to set a realistic upper bound (e.g., body weight). The goal is to provide the risk manager with an "unbiased estimate of risk." One expert reiterated that subjective judgments are inherent in the risk assessment process. In the case of truncating data, such judgments must be explained clearly and justified to the risk manager. In contrast to truncation, one expert reminded the group that the risk manager decides upon what percentile of the tail is of interest. Because situations arise in which the risk manager may be looking for 90th to 99th percentile values, the assessor must know how to approach the problem and, ultimately, must clearly communicate the approach and the possible large uncertainties.

Scenarios. The group discussed approaches for evaluating the *high ends of distributions* (e.g., the treatment blips mentioned previously or the pica child). Should the strategy for

assessing overall risks include high end or unusual behavior? Several experts felt that including extreme values in the overall distribution was not justified and suggested that the upper bounds in these cases be considered "scenarios." As with upper bounds, one expert noted that low end values also need special attention in some cases (e.g., air exchange in a tight house).

Generalized distributions versus mixtures. Expert opinion differed regarding the issue of generalized versus mixture distributions. One expert was troubled by the notion of a mixture distribution. He would rather use a more sophisticated generalized distribution. Another expert provided an example of radon, stating that it is likely a mixture of Normal distributions, not a Lognormal distribution. Therefore, treatment of mixtures might be a reasonable approach. Otherwise, assessors risk grossly underestimating risk in concentrated areas by thinking they know the parametric form of the underlying distribution.

  The same expert noted that the issue of mixtures highlights the importance of having some theoretical basis for applying available techniques (e.g., possible Bayesian methods). Another expert stated that he could justify using distributions that are mixtures, because in reality many data sets are inherently mixtures.

Truncation of distributions. Mixed opinions were voiced on this issue. One expert noted that assessors can *extend a distribution to a plausible upper bound* (e.g., assessors can predict air exchange rates because they know at a certain point they will not go higher). Another expert noted that truncating the distribution by 2 or 3 standard deviations is not uncommon because, for example, the assessors simply do not want to generate 1,500-pound people. One individual questioned, however, whether truncating a Lognormal distribution invalidates calling the distribution Lognormal. Another commented on instances in which truncating the distribution may be problematic. For example, some relevant data may be rejected. Also, the need to truncate suggests that the fit is very poor. The only reason to truncate, in his opinion, is if one is concerned about getting a zero or negative value, or perhaps an extremely high outlier value. One expert noted that truncation clearly has a role, especially when a strong scientific or engineering basis can be demonstrated.

When should neither an EDF nor PDF be used? Neither an EDF nor a PDF may be useful/appropriate when large extrapolations are needed or when the assessor is uncomfortable with extrapolation beyond the available data points. In these cases, scenario analyses may come into play.

In their final discussions on EDF/PDF, the group widely encouraged *visual or graphical representation* of data. Additional thoughts on visually plotting the data are presented in the following discussions of goodness of fit.

6.2 GOODNESS-OF-FIT (GoF)

On what basis should it be decided whether a data set is adequately represented by a fitted parametric distribution?

The final workshop discussions related to the appropriateness of using available GoF test statistics in evaluating how well a data set is represented by a fitted distribution. Experts were asked to consider what options are best suited and how one chooses among multiple tests that may provide different answers. The following highlights the major points of these discussions.

- *Interpreting poor fit.* GoF in the middle of the distribution is not as important as that of the tails (upper and lower percentiles). Poor fit may be due to outliers at the other end of the distribution. If there are even only a few outliers, GoF tests may provide the wrong answer.

- *Graphical representation* of data is key to evaluating goodness or quality of fit. Unanimously, the experts agreed that using probability plots (e.g., EDF, QQ plots, PP plots) or other visual techniques in evaluating goodness of fit is an acceptable and recommended approach. In fact, the group felt that graphical methods should always be used. Generally, it is easier to judge the quality of fit using probability plots that compare data to a straight line. There may be cases in which a fit is rejected by a particular GoF test but appears reasonable when using visual techniques.

  The group supported the idea that GoF tests should not be the only consideration in fitting a distribution to data. Decisions can be made based on visual inspection of the data. It was noted that graphical presentations help to show quirks in the data (e.g., mixture distributions). It was also recommended that the assessor seek the consensus of a few trained individuals when interpreting data plots (as is done in the medical community when visually inspecting X-rays or CAT scans).

- *What is the significance of failing a weak test such as chi-square? Can we justify using data that fail a GoF test?* GoF tests may be sensitive to imperfections in the fit that are not important to the assessor or decision maker. The group therefore agreed that the fitted distribution can be used especially if the failure of the test is due to some part of the distribution that does not matter to the analysis (e.g., the lower end of the distribution). The reason the test failed, however, must be explained by the assessor. Failing a chi-square test is not problematic if the lower end of the distribution is the reason for the failure. One expert questioned whether the assessor could defend (in court) a failed statistical test. Another expert responded indicating that a graphical presentation might be used to defend use of the data, showing, for example, that the poor fit was a result of data set size, not chance.

- Considerations for risk assessors when GoF tests are used.

— The evaluation of distributions is an estimation process (e.g., PDFs). Using a systematic testing approach based on the straight line null hypothesis may be problematic.

— R^2 is a poor way to assess GoF.

— The appropriate loss function must be identified.

— The significance level must be determined before the data are analyzed. Otherwise, it is meaningless. It is a risk management decision. The risk assessor and risk manager must speak early in the process. The risk manager must understand the significance level and its application.

Should GoF tests be used for parameter estimation (e.g., objective function is to minimize the one-tail Anderson-Darling)? A degree of freedom correction is needed before the analysis is run. The basis for the fit must be clearly defined—are the objective and loss functions appropriate?

"Maximum likelihood estimation (MLE)" is a well-established statistical tool and provides a relatively easy path for separating variability from uncertainty.

The adequacy of Crystal Ball®'s curve-fitting capabilities was questioned. One of the experts explained that it runs three tests, then ranks them. If the assessor takes this one step further by calculating percentiles and setting up plots, it is an adequate tool.

The Office of Water collects large data sets. Some of the office's efforts might provide some useful lessons into interpreting data in the context of this workshop.

What do we do if only summary statistics are available? Summary statistics are often all that are available for certain data sets. The group agreed that MLE can be used to estimate distribution parameters from summary data. In addition, one expert noted that probability plots are somewhat useful for evaluating percentile data. Probability plots enable assessors to evaluate the slope (standard deviation) and the intercept (mean). Confidence intervals cannot be examined and uncertainty cannot be separated from variability.

In summary, the group identified possible weaknesses associated with using statistical GoF tests in the context described above. The experts agreed unanimously that graphical/visual techniques to evaluate how well data fit a given distribution (alone or in combination with GoF techniques) may be more useful than using GoF techniques alone.

6.3 SUMMARY OF EDF/PDF AND GoF DISCUSSIONS

The experts agreed, in general, that the choice of an EDF versus a PDF is a matter of personal preference. The group recommended, therefore, that no rigid guidance be developed requiring one or the other in a particular situation. The decision on which distribution function to use is dependent on several factors, including the number of data points, the outcome of interest, and how interested the assessor is in

the tails of the distribution. Varied opinions were voiced on the use of mixture distributions and the appropriateness of truncating distributions. The use of scenario analysis was suggested as an alternative to probabilistic analysis when a particular input cannot be assigned a probability distribution or when estimating the tails of an important input distribution may be difficult.

Regarding GoF, the group fully agreed that visualization/graphic representation of both the data and the fitted distribution is the most appropriate and useful approach for ascertaining adequacy of fit. In general, the group agreed that conventional GoF tests have significant shortcomings and should not be the primary method for determining adequacy of fit.

SECTION SEVEN

OBSERVER COMMENTS

This section presents observers' comments and questions during the workshop, as well as responses from the experts participating in the workshop.

DAY ONE: Tuesday, April 21, 1998

Comment 1
Helen Chernoff, TAMS Consultants

Helen Chernoff said that, with the release of the new policy, users are interested in guidance on how to apply the information on data representativeness and other issues related to probabilistic risk assessment. She had believed that the workshop would focus more on application, rather than just on the background issues of probabilistic assessments. What methods could be used to adjust data and improve data representativeness (e.g., the difference between past and current data usage)?

Response

The workshop chair noted that adjustment discussions during the second day of the workshop start to explore available methods. One expert stated that, based on his impression, the workshop was designed to gather input from experts in the field of risk assessment and probabilistic techniques. He noted that EPA's policy on probabilistic analysis emerged only after the 1996 workshop on Monte Carlo analysis. Similarly, EPA will use the information from this workshop to help build future guidance on probabilistic techniques, but EPA will not release specific guidance immediately (there may be an approximate two-year lag).

The chair noted that assessors may want to know when they can/should implement alternate approaches. He pointed out that the representativeness issue is not specific to probabilistic assessment. It applies to all assessments. Since EPA released its May 1997 policy on Monte Carlo analysis, representativeness has been emphasized more, especially in exposure factor and distribution evaluations. He noted, however, that data quality/representativeness is equally important when considering a point estimate. However, it may not be as important if a point estimate is based on central tendency instead of an upper percentile where there may be fewer data. Another agreed that the representativeness issue is more important for probabilistic risk assessment than deterministic risk assessment (especially a point estimate based on central tendency).

Comment 2
Emran Dawoud, Human Health Risk Assessor, Oak Ridge National Laboratory

Mr. Dawoud commented that the representativeness question should reflect whether additional data must be collected. He noted that the investment (cost/benefit) should be considered. From a risk assessment point of view, one must know how more data will affect the type or cost of remedial activity. In his opinion, if representativeness does not change the type or cost of remedial activity, further data collection is unwarranted.

Mr. Dawoud also commented that the risk model has three components: source, exposure, and dose-response. Has the sensitivity of exposure component been measured relative to the sensitivity of the other two components? He noted the importance of the sensitivity of the source term, especially if fate and transport are involved.

Mr. Dawoud briefly noted that, in practice, a Lognormal distribution is being fit with only a few samples. Uncertainty of the source term in these cases is not quantified or incorporated into risk predictions. Even if standard deviation is noted, the contribution to final risk prediction is not considered. Mr. Dawoud noted that the workshop discussions on the distribution around exposure parameters seem to be less important than variation around the source term. Likewise, he noted the uncertainties associated with the dose-response assessment as well (e.g., applying uncertainty factors of 10, 100, etc.).

Response

One participant noted that representativeness involves more than collecting more data. Evaluating representativeness is often about choosing from several data sets. He agreed that additional data are collected depending on how the collection efforts may affect the bottom line assessment answer. He noted that if input does not affect output, then its distribution need not be described.

Relative to Mr. Dawoud's second point, it was noted that source term evaluation is part of exposure assessment. While exposure factors (e.g, soil ingestion and exposure duration) affect the risk assessment, one expert emphasized that the most important driving "factor" is the source term. As for dose-response, the industry is just beginning to explore how to quantify variability and uncertainty.

The workshop chair noted that methodologically, exposure and source terms are not markedly different. The source term has representativeness issues. There are ways to distinguish between variability and uncertainty in the variability estimate.

Lastly, more than one expert agreed that the prediction of risk for noncancer and cancer endpoints (based on the reference dose [RfD] and cancer slope factor [CSF], respectively) is very uncertain. The methods discussed during this workshop cannot be directly applied to RfDs and CSFs, but they could be used on other toxicologic data. More research is needed in this area.

Comment 3
Ed Garvey, TAMS Consultants

Mr. Garvey questioned whether examining factors of 2 or 3 on the exposure side is worthwhile, given the level of uncertainty on the source or dose-response term, which can be orders of magnitude.

Response

It was an EPA policy choice to examine distributions looking first at exposure parameters, according to one EPA panelist. He also reiterated that the evaluation of exposure includes the source term (i.e., exposure = concentration x uptake/averaging time). One person noted that it was time to "step up" on quantifying toxicity uncertainty. Exposure issues have been driven primarily by engineering approaches (e.g., the Gaussian plume model), toxicity has historically been driven by toxicologists and statisticians and are more data oriented.

It was noted that, realistically, probabilistic risk assessments will be seen only when money is available to support the extra effort. Otherwise, 95% UCL concentrations and point estimates will continue to be used. Knowing that probabilistic techniques will enable better evaluations of variability and uncertainty, risk assessors must be explicitly encouraged to perform probabilistic assessments. We must accept that the existing approach to toxicity assessment, while lacking somewhat in scientific integrity, is the only option at present.

Comment 4
Emran Dawoud, Human Health Risk Assessor, Oak Ridge National Laboratory

Mr. Dawoud asked whether uncertainty analysis should be performed to evaluate fate and transport related estimates.

Response

One expert stressed that whenever direct measurements are not available, variability must be assessed. He commented that EPA's Probabilistic Risk Assessment Work Group is preparing two chapters for Risk Assessment Guidance for Superfund (RAGS): one on source term variability and another on time-dependent considerations of the source term.

Comment 5
Zubair Saleem, Office of Solid Waste, U.S. EPA

Mr. Saleem stated that he would like to reinforce certain workshop discussions. He commented that any guidance on probabilistic assessments should not be too rigid. Guidance should clearly state that methodology is evolving and may be revised. Also, guidance users should be encouraged to collect additional data.

Response

The workshop chair recognized Mr. Saleem's comment, but noted that the experts participating in the workshop can only provide input and advice on methods, and is not in a position to recommend specific guidelines to EPA.

DAY TWO: Wednesday, April 22, 1998

Comment 1
Lawrence Myers, Research Triangle Institute

Mr. Myers offered a word of caution regarding GoF tests. He agrees that many options do not work well but he stated that in an adversarial situation (e.g., a court room) he would rather be defending data distributions based on a quantitative model instead of a graphical representation.

Mr. Myers noted that the problem with goodness of fit is the tightness of the null hypothesis (i.e., it specifies that the true model is exactly a member of the particular class being examined). Mr. Myers cited Hodges and Layman (1950s) who generalized chi-square in a way that may be meaningful to the issues

discussed in this workshop. Specifically, because exact conformity is not expected, a more appropriate null hypothesis would be that the true distribution is "sufficiently close" to the family being examined.

Response

One expert reiterated that when a PDF is fitted, it is recognizably an approximation and therefore makes application of standard GoF statistics difficult. Another expressed concern that practitioners could go on a "fishing expedition," especially in an adversarial situation, to find a GoF test that gives the right answer. He did not feel this is the message we want to be giving practitioners. A third expert noted a definite trend in the scientific community away from GoF tests and towards visualization.

SECTION EIGHT

REFERENCES

Buck, R.J., K.A. Hammerstrom, and P.B. Ryan, 1995. Estimating Long-Term Exposures from Short-term Measurements. *Journal of Exposure Analysis and Environmental Epidemiology*, Vol. 5, No. 3, pp. 359-373.

Efron, B. and R.J. Tibshirani, 1993. An Introduction to the Bootstrap. Chapman and Hall. New York.

Frey, H.C. and D.E. Burmaster, "Methods for Characterizing Variability and Uncertainty: Comparison of Bootstrap Simulation and Likelihood-Based Approaches," *Risk Analysis* (Accepted 1998).

RTI, 1998. *Development of Statistical Distributions for Exposure Factors*. Final Report. Prepared by Research Triangle Institute. U.S. EPA Contract 68D40091, Work Assignment 97-12. March 18, 1998.

U.S. Environmental Protection Agency, 1996a. Office of Research and Development, National Center for Environmental Assessment. *Exposure Factors Handbook*, SAB Review Draft (EPA/600/P-95/002Ba).

U.S. Environmental Protection Agency, 1996b. *Summary Report for the Workshop on Monte Carlo Analysis*. EPA/630/R096/010. September 1996.

U.S. Environmental Protection Agency, 1997a. *Guiding Principles for Monte Carlo Analysis*. EPA/630/R-97/001. March 1997.

U.S. Environmental Protection Agency, 1997b. Policy for Use of Probabilistic Analysis in Risk Assessment at the U.S. Environmental Protection Agency. May 15, 1997.

Wallace, L.A., N. Duan, and R. Ziegenfus, 1994. Can Long-term Exposure Distributions Be Predicted from Short-term Measurements? *Risk Analysis*, Vol. 14, No. 1, pp. 75-85.