United States
Environmental Protection
Agency

Office of Policy, Planning
and Evaluation
Washington, DC 20460

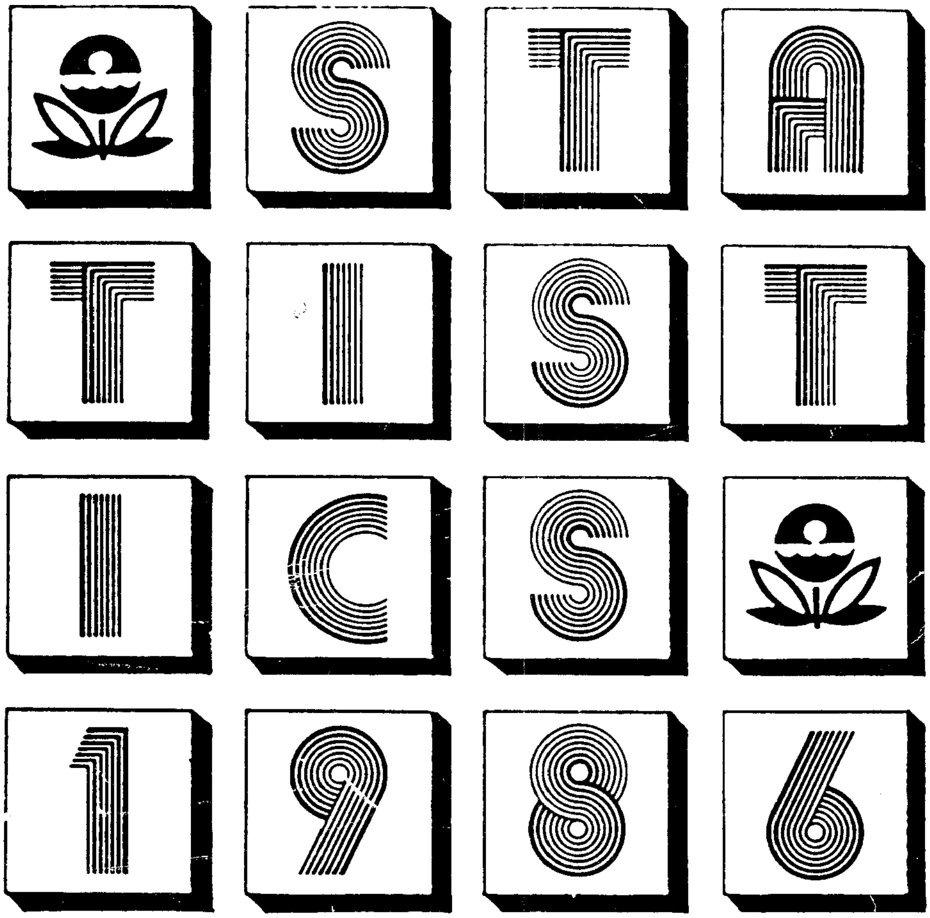
EPA-230-12-87-032

Statistical Policy Branch

EPA

ASA/EPA Conferences on Interpretation of Environmental Data

II. Statistical Issues in Combining Environmental Studies October 1–2, 1986



DISCLAIMER

This document has not undergone final review within EPA and should not be used to infer EPA approval of the views expressed.

PREFACE

This volume is a compendium of the papers and commentaries that were presented at the second of a series of conferences on interpretation of environmental data conducted by the American Statistical Association and the U.S. Environmental Protection Agency's Statistical Policy Branch of the Office of Standards and Regulations/Office of Policy, Planning, and Evaluation.

The purpose of these conferences is to provide a forum in which professionals from the academic, private, and public sectors can exchange ideas on statistical problems that confront EPA in its charge to protect the public and the environment through regulation of toxic exposures. They provide a unique opportunity for Agency statisticians and scientists to interact with their counterparts in the private sector.

The theme of the conference, "Statistical Issues in Combining Environmental Studies," is particularly appropriate because policy formulation rarely depends upon a single study. At any rate, the conclusions from various studies are often seemingly contradictory or the evidence from any single study is not clear-cut. No matter how inconclusive the evidence may be, it is still necessary to formulate policies. Recently, great strides have been made in the formal statistical combination of research information. A new term, "meta-analysis," has appeared, most often in medicine and social science research reports. This ASA/EPA research conference was held to make environmental statisticians and scientists aware of these new techniques and to examine the applicability of the methodology to environmental studies.

The holding of a research conference and preparation of papers for publication requires the efforts of many people. Gratitude is expressed to the ASA Committee on Statistics and the Environment which was instrumental in developing this series of conferences. In addition, appreciation is given to Dr. Kinley Larntz, University of Minnesota, for his work in assembling and coordinating the presentations for this conference. Thanks are also owed to members of the ASA staff and, particularly, Ede Denenberg and Mary Esther Barnes, who supported the entire effort. Although there was no provision for a formal peer review, thanks are also due to the reviewers who assessed the articles for their scientific merit and raised questions which were submitted to the authors for their consideration.

The views presented in this conference are those of individual writers and should not be construed as reflecting the official position of any agency or organization.

Following the first conference on "Current Assessment of Combined Toxicant Effects," in May 1986, the second conference on "Statistical Issues in Combining Environmental Studies" was held in October 1986. Two additional conferences were held: "Sampling and Site Selection for Environmental Studies" in May 1987 and "Compliance Sampling" in October 1987. From these two conferences, proceedings volumes will also be published.

Emanuel Landau, Editor
American Public Health Association

Dorothy G. Wellington, Co-Editor
U.S. Environmental Protection Agency

SUMMARY

The first set of papers and discussion in the volume begins with an exciting paper by David M. Eddy introducing a Bayesian method for evaluating and summarizing evidence from various sources. The sources can include empirical studies and expert judgments. Graphical methods are provided to display the conclusions from the research synthesis. One of the highlights of the conference was an interactive computing demonstration of the techniques given by Eddy and Vic Hasselblad. Robert L. Wolpert's paper discusses the critical issue of selecting the appropriate scale of measurement for combining evidence. Following the two papers is discussion by David A. Lane. Lane offers strong support for the Bayesian viewpoints of Eddy and Wolpert. He also reminds us of potential difficulties in implementation of the methods.

Larry Hedges, a major contributor to the meta-analysis literature, presents a clear picture of the issues in combining studies. In contrast to Eddy, Wolpert, and Lane, Hedges adopts the frequentist viewpoint presenting combined significance tests an confidence limits. Hedges carefully presents methods and points out their possible limitations. In discussion of the Hedges paper, James M. Landwehr point out that combining studies should be considered within the usual framework of statistical applications. Landwehr then takes us through the steps of standard analysis, illustrating the special aspects of meta-analysis.

In the next paper, Thomas B. Feagans presents his viewpoint on probabilistic assessments. Feagans bases his methods on the fundamental axioms of probability. Interesting discussions are given by Harvey M. Richmond, Anthony D. Thrall, and Miley W. Merkhofer.

The final paper, given by G.P. Patil, G.J. Babu, M.T. Boswell, K. Chatterjee, E. Linder, and C. Taillie, presents several case studies of combining data in the environmental area, specifically in marine fisheries management. Lloyd L. Lininger gives a discussion raising fundamental questions important to any problem of combining studies.

Kinley Larntz
University of Minnesota

INDEX OF AUTHORS

TABLE OF CONTENTS

CONFIDENCE PROFILES: A BAYESIAN METHOD FOR ASSESSING HEALTH TECHNOLOGIES

David M. Eddy, M.D., Ph.D.*
J. Alexander McMahon Professor of
Health Policy and Management,
Director
Center for Health Policy Research and Education
Duke University
Durham, North Carolina

INTRODUCTION

The first step in the assessment of a health technology is to evaluate the existing evidence to estimate how the technology affects the magnitude or probability of important health outcomes—its benefits and harms. These estimates form the basis for the subsequent steps of an assessment: comparison of benefits and harms, estimation of overall benefit, calculation of marginal returns, and design of a policy.

At present, for the great majority of health technologies there is no explicit quantitative estimation of the technology's effects on health outcomes. Current clinical and administrative decisions are typically based on a qualitative subjective judgment that a technology's benefits outweigh its harms. However, the rising cost of health care, increasing competition, concern over wide variations in practice patterns, increasing malpractice claims, and a variety of other forces all create pressure for quantitative estimates of a technology's effects, and therefore for quantitative assessment methods.

Estimating the effects of a technology on health outcomes is complicated by several factors. Ideally, for each technology and each outcome, there would be several well designed controlled trials that provide direct evidence of how the technology affects each outcome. Unfortunately, this ideal is rarely achieved. The empirical evidence is rarely complete. What evidence is available usually comes from many different sources, including randomized controlled trials (RCTs), nonrandomized controlled trials, uncontrolled clinical series, case-control studies, cross-sectional studies, case reports, longitudinal studies, and animal experiments. Even anecdotes, theories, testimonies, and analogies play a role in many assessments. Each piece of evidence can be subject to a variety of biases and other factors that affect their internal validity, comparability, and applicability to a particular assessment (external validity). Much of the available evidence does not deal with outcomes that are important to patients (e.g., death), but with intermediate outcomes (e.g., cholesterol level), or performance indicators (e.g., the sensitivity of a diagnostic test). Finally, even pieces of evidence that are complete and have the same design can be inconclusive (e.g., not statistically significant) or give inconsistent results. Because of these complexities, the process for synthesizing evidence tends to be highly subjective—which leave them vulnerable to oversimplification, errors in reasoning, wishful thinking, and self-interest.

This paper introduces a Bayesian method for synthesizing the available evidence—from both empirical studies and expert judgments—to estimate the effect of a health technology on health outcomes. Called the *Confidence Profile Method*, it can be used to evaluate evidence from different types of empirical studies, adjust individual pieces of evidence for biases to internal and external validity, combine evidence from different studies (not necessarily with the same designs), and incorporate focused subjective judgments, to derive a probability distribution for the effect of a health technology on health outcomes.[1] Because the probability distribution explicitly incorporates subjective judgments, it is called a *Confidence Profile*. The Profiles for each outcome can then form the basis for adjustments for risk aversion, comparison of benefits and harms, and other steps of a technology assessment.

This paper gives the basic formulas of the method, and illustrates its use with an analysis of the effect of a thrombolytic agent—tissue-type plasminogen activator—on one-year survival from heart attacks.

DEFINITIONS

The term *health technology* is used very broadly to include any intervention that might affect a health outcome. Examples include health education, diagnostic tests, treatments, rehabilitation programs, pain control, and psychotherapy. A *health outcome* is an outcome of a disease or injury that people can experience and care about. Examples are life and death, pain, disfigurement, disability, anxiety, and range of motion of a limb. It is important to distinguish health outcomes from *intermediate outcomes*, which are markers of biological changes that might indicate or affect the probability or magnitude of health outcomes. Examples are blood pressure, serum cholesterol, intraocular pressure, and the reperfusion of a coronary artery after treatment of a heart attack.[2]

The objective of an assessment is to estimate the technology's effect on health outcomes. To accomplish this, we use "chains" that connect the performance of the technology to the health outcome. If there is *direct* evidence that directly relates performance of the technology to the occurrence of the health outcome, the chain has a single link. In other cases the evidence is *indirect*, with one body of evidence relating the performance of the technology to one or more intermediate

*I thank Vic Hasselblad, Greg Critchfield, Dick Smallwood, and Bob Winkler for many helpful suggestions. I especially thank Robert Wolpert for his contributions to this work and collaboration in extending it.

outcomes (or followup actions—see below), and other evidence relating the intermediate outcome(s) to the health outcome. For example (see illustration below), to assess the effect of changing dietary cholesterol on heart attack rates, a two-link chain might be used; the first link would evaluate evidence about the effect of diet (the technology) on serum cholesterol (the intermediate outcome); the second link would evaluate evidence that reducing serum cholesterol reduces heart attack rates (the health outcome). When multiple-link chains are used, care must be taken to examine the accuracy of the intermediate outcome as an indicator for the health outcome, and any independent effects of the technology on the health outcome (not mediated through the intermediate outcome). These issues will be discussed in detail below.

Change diet——> lower serum cholesterol——>
prevent heart attacks

Followup actions are important in the evaluation of diagnostic or screen technologies, where the technology's purpose is to provide information, which in turn can affect health outcomes only if it changes a followup action (e.g., changes treatment). For example, to evaluate screening for ocular hypertension, a three-link chain would be constructed to (1) relate the use of the screening test (e.g., tonometry) to detection of high intraocular pressure (an intermediate outcome), relate detection of elevated pressure to a decision to treat (a followup action), and (3) relate the treatment to a decrease in the chance of blindness (a health outcome).

Frequently there are features of the population, disease, technology, provider or setting that can alter the effect of a technology on health outcomes. Examples are the relative risk of a disease in a population to be screened, the sensitivity or specificity of a diagnostic test, the dose or frequency of a drug, the experience of a practitioner, and the adherence of a patient to a treatment. The Confidence Profile Method treats these features as *parameters*. By performing an assessment as a function of various parameters, the assessment's results can be tailored to a variety of circumstances.

MEASURING A TECHNOLOGY'S EFFECT

To estimate a technology's effect on health outcomes, a suitable measure must be chosen for each outcome, and an estimate made of how the technology (compared with a designated control) changes the outcome, according to the chosen measure. Quantitative measures provide the least ambiguous way to describe, and the most powerful way to calculate, a technology's effect.

Given a quantitative measure for a health outcome, the effect of a technology can be defined in several different ways. For dichotomous health outcomes, an obvious measure of effect is the change in the probability of the health outcome. For example, if the chance of the health outcome (e.g., death within one year) without the technology is 0.8, and the chance of the health outcome with the technology is 0.4, the effect of the technology by this measure is -0.4. (The technology caused the chance of death to be 40% less than without the technology.) For health outcomes that can take one of several values (e.g., mild, moderate or severe pain, or a discrete-valued health status measure), the technology's effect can be defined as the shift in probabilities of the different outcomes. For continuous-valued outcomes (e.g., weight, IQ, or a continuous-valued health status measure), the technology's effect can be measured as the change in the magnitude of the health outcome. Other measures of a technology's effect are possible, such as a change in the odds-ratio, or the percent change in probability or magnitude of an outcome. In each case, uncertainty about the technology's effect can be described as a distribution for the effect, on the chosen measure.

OVERVIEW OF THE CONFIDENCE PROFILE METHOD

Steps. The Confidence Profile Method is applied in five basic steps. This section outlines the steps and the general form of some of the formulas. Examples of specific formulas are given in a later section.

1. Define the technology, the control with which it will be compared (the "designated control"), the circumstances in which it will be applied (the "circumstances of interest"), and the health outcomes it affects. A separate assessment should be performed for each health outcome.

2. For each health outcome, describe chain(s) that relate the performance of the technology (compared with the designated control) to the occurrence of the health outcome. These chains should be based on the available evidence and knowledge of the pathophysiology and management oi the health problem. The chains should be created so that each piece of evidence applies to one and only one chain. Each chain will be analyzed separately, and the results combined in a later step (step 4).

3. For each chain, derive a probability distribution for the effect of the technology on the health outcome, as indicated by the evidence for that chain. This is accomplished by examining the evidence for each link of the chain, one link at a time, deriving a probability distribution for the link (step 3a), and then combining the probability distributions across the chain (step 3c).

   a. The derivation of a probability distribution for a link, that describes our knowledge about the true effect of the action on the outcome for that link, is accomplished by first deriving for each link a likelihood function L(•) for the likelihood of the observed results of the evidence for the link as a function of the possible values of the true effect.[3] To simplify the discussion, consider a single-link chain (direct evidence) and denote as ϵ the true (but unknown) effect of the technology on the health outcome. Denote the observed results of an experiment or other source of evidence as X_{ij}, where the subscripts denote the i^{th} piece of evidence for the j^{th} chain. Where there is no ambiguity about which chain is being considered, the second subscript will be dropped. (Below, the collection of evidence for the j^{th} chain will be denoted $X_{\cdot j}$, and the total body of evidence for all chains will be denoted

$X_{..}$). The likelihood function we want to derive for the link, based on, say, *n* pieces of evidence, is $L(\varepsilon|X_1, X_2, ..., X_n)$.

i. To derive this likelihood function for the link, examine each independent piece of evidence for the link one by one and derive a function for the likelihood of the observed result (X_i) as a function of the possible values of the true effect of the technology (ε). Denote this likelihood function for the i^{th} piece of evidence as $L_i(\varepsilon|X_i)$. The form of the likelihood function will depend on the type of evidence. The likelihood function for an RCT will be given in the next section.

ii. Sometimes the result of a particular piece of evidence is influenced by factors that affect internal validity (e.g., patient selection bias, errors in measurement of outcomes, crossover of patients between treated and control groups), external validity (e.g., differences between the circumstances of a trial relating to the patients, technology, or providers, compared with the circumstances of interest in a particular assessment). Because of this, the formulas for deriving likelihood functions for individual pieces of evidence contain variables to adjust for these factors. The requirement is that, when the appropriate adjustments are made, the likelihood function for each piece of evidence should describe the likelihood of the *observed* results of the study *in the circumstances of the study,* as a function of the true effect of the technology *in the circumstances of interest.* Specific examples of likelihood functions that contain adjustments will be given in the next section.

iii. Calculate a joint likelihood function for the observed results of all pieces of evidence, as a function of the (unknown) true effect of the technology in the circumstances of interest, by multiplying the likelihood functions of the individual pieces of evidence (possibly adjusted for biases to internal and external validity).[4]

$$L(\varepsilon|X_1, X_2, ..., X_n) = L_1(\varepsilon|X_1)\, L_2(\varepsilon|X_2) \cdots L_n(\varepsilon|X_n) \qquad (1)$$

iv. Derive a probability distribution for the effect of the action on the outcome (for that link), using the continuous form of Bayes' formula. Denote this (posterior) probability distribution by $\pi(\varepsilon|X_.)$. Thus

$$\pi(\varepsilon|X_1, X_2, ..., X_n) = k\, L(\varepsilon|X_1, X_2, ..., X_n)\, \pi(\varepsilon) \qquad (2)$$

where $\pi(\varepsilon)$ is a noninformative prior distribution for ε, and k is a normalizing constant. The choice of a suitable noninformative prior distributions is discussed by Jeffreys (1961) and Bernardo (1979) in a general setting, and by Wolpert and Eddy (1986) in the context of Confidence Profiles.

b. If the evidence is direct (a single-link chain), the posterior probability distribution derived in the previous step is the Confidence Profile for the effect of the technology (skip to step 4). If the evidence is indirect (a multiple-link chain), repeat step 3a to derive probability distributions for each link and proceed to step 3c.

c. Combine the probability distributions for each link to derive a probability distribution for the entire chain—the effect of the technology on the health outcome. If the occurrence of an outcome for a particular link is determined solely by the action for that link, and not affected by any preceding actions, (e.g., if the technology has no independent effect on the health outcome not mediated through the intermediate outcome),[5] then the probability distributions for the links are combined by an operation analogous to multiplication of two random variables. Specifically, let $\pi_{t\omega}(\varepsilon_{t\omega})$ be the distribution for the effect of the technology (t) on a dichotomous intermediate outcome (ω), let $\pi_{\omega h}(\varepsilon_{\omega h})$ be the distribution for the effect of the intermediate outcome on a dichotomous health outcome (h), where in both cases the effect is measured as the difference in the probability of the outcome caused by the action (see footnote 3). Then the distribution for the effect of the technology on the health outcome ($\pi_{th}(\varepsilon_{th})$) is given by

$$\pi_{th}(\varepsilon_{th}) = \int_{-\infty}^{\infty} \frac{1}{|\varepsilon_{t\omega}|}\, \pi_{t\omega}(\varepsilon_{t\omega})\, \pi_{\omega h}\,(\varepsilon_{th}/\varepsilon_{t\omega})\, d\varepsilon_{t\omega} \qquad (3)$$

If the occurrence of an outcome is not determined solely by the action for that link, but is influenced by an action in a preceding link, the formula for combining probability distributions across links of a chain must include correction factors. The equation with correction factors is given below (Eq. [25]).

The result of this step is the probability distribution, or *Confidence Profile,* for the effect of the technology (compared with the designated control) on the health outcome, for a particular chain.

4. If there are two or more chains, combine their separate probability distributions to derive a single probability distribution that incorporates the evidence in all the chains. Let $X_{.i}$ be the evidence for the i^{th} chain, let $\pi_i(\varepsilon|X_{.i})$ be the probability distribution for the effect of the

technology, based on the evidence in the i^{th} chain, and let $\pi(\varepsilon)$ be the (noninformative) prior for the technology's effect. The formula is (4)

$$\pi(\varepsilon|X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot n}) = k\, \pi_1(\varepsilon|X_{\cdot 1})\, \pi_2(\varepsilon|X_{\cdot 2}) \cdots \pi_n(\varepsilon|X_{\cdot n}) \cdot \left[\frac{1}{\pi(\varepsilon)}\right]^{n-1}$$

where k is a normalizing constant and n is the number of chains. This equation assumes the posterior distributions for each chain are independent in the sense that no piece of evidence is used in more than one chain. Equation (4) will be derived below (Eq.[33]).

5. Sometimes a body of evidence will compare the technology (*T*) with a control (C^*) that is different from the designated control (*C*). When this occurs, the effect of the technology compared with the designated control (call this ε_{tc}) can be found as follows:

   a. use steps 1–4 to derive a probability distribution for the effect of *T* compared with C^* (call this ε_{tc^*}),

   b. similarly, derive a probability distribution for the effect of C^* versus the designated control *C* (call this ε_{c^*c}) [6] and

   c. calculate the probability distribution for the effect of the technology compared with the designated control by convolving the probability distributions outlined in the previous two steps. Specifically, let $\pi_{tc^*}(\varepsilon_{tc^*})$ be the distribution for the effect of the technology compared with the control C^*, let $\pi_{tc}(\varepsilon_{tc})$ be the distribution for the effect of the technology compared with the designated control *C*, and let $\pi_{c^*c}(\varepsilon_{c^*c})$ be the distribution for the effect on the health outcome of the control C^* compared with the designated control *C*. Then

$$\pi_{tc}(\varepsilon_{tc}) = \int_{-\infty}^{\infty} \pi_{tc^*}(\varepsilon_{tc^*})\, \pi_{c^*c}(\varepsilon_{tc}-\varepsilon_{tc^*})\, d\varepsilon_{tc^*} \qquad (5)$$

Use of Subjective Judgments. When applying the formulas of the Confidence Profile Method, empirical evidence is used to the greatest extent possible to estimate the necessary variables. However, whenever the available empirical evidence is incomplete, focused subjective judgments must be used to complete an assessment.[7] An important feature of the Confidence Profile Method is that whenever subjective judgments are used, uncertainty about any variable being estimated can be incorporated in the analysis by describing a probability distribution for the variable (instead of using a point estimate), and integrating over the variable. No additional subjective "weighing" of individual pieces of evidence is required; the "weight" of each piece of evidence (along with any adjustments) is automatically captured in the likelihood functions and therefore in the Confidence Profiles calculated from them. This feature will be illustrated below. Because the Confidence Profile automatically encodes this uncertainty about the variables in the formulas, as well as the uncertainty due to the random sampling that affects empirical observations, there is no need to perform sensitivity analysis for such factors.[8]

ILLUSTRATION

The Confidence Profile Method will be illustrated with formulas for evidence involving RCTs with dichotomous intermediate outcomes and health outcomes, where the technology's effect is measured as the difference it causes in the probability of the health outcome. Use of the formulas will be illustrated with an assessment of the effect on one-year survival of a thrombolytic agent (tissue-type plasminogen activator) used to treat heart attacks. The Method currently includes formulas for other types of experimental designs (e.g., nonrandomized controlled trials, clinical series, case-control studies, and some cross-sectional designs); categorical and continuous-valued intermediate outcomes and health outcomes; and other measures of a technology's effect (e.g., change in odds-ratio, and percent change in a rate). These are described elsewhere (Eddy and Wolpert 1986).

Background. Tissue-type plasminogen activator (t-PA) is one of several thrombolytic agents used to dissolve (lyse) blood clots (thrombi) in coronary arteries after heart attacks, with the intention of restoring blood flow through the coronary artery (reperfusion), and thereby increasing the chance of survival. There are conflicting policies about the use of t-PA (and about payment for it by third-party payers). The conflicts reflect the complexity of the available evidence. The main problem is that there is no single RCT that compares the effect of t-PA with conventional care or any other thrombolytic agents on long-term (one-year) survival. The available studies of t-PA (see Table I) involve intermediate outcomes (e.g., perfusion and reperfusion), short-term outcomes (in-hospital mortality), and different controls (placebo, conventional care, and intravenous streptokinase). In addition to the studies described in Table I, a large number of studies have examined other thrombolytic agents—intravenous streptokinase (IV SK), intracoronary streptokinase (IC SK), and urokinase (UK) (Yusuf et al 1985). While they do not provide direct evidence about t-PA, they contain information to compare the various controls used in studies involving t-PA.

Likelihood Function for an RCT. The likelihood function for an RCT is derived from the binomial distribution. Designate the occurrence of the health outcome a "success" (*s*), the nonoccurrence of the outcome a "failure" (*f*), and the true probability of a success in the treated and control groups as p_1 and p_0, respectively. Let the number of people in the treated and control groups be n_1 and n_0, the observed number of successes in each group be s_1 and s_0, and the observed number of failures in each group be f_1 and f_0 ($s_i + f_i = n_i$). The joint likelihood function for p_0 and p_i, given observed values of s_0, f_0, s_i, and f_1 is

$$L(p_0, p_1|s_0, f_0, s_1, f_1) = (p_0)^{s_0}\,(1-p_0)^{f_0}\,(p_1)^{s_1}\,(1-p_1)^{f_1} \qquad (6)$$

Designate the "effect" of a technology as ε, which in this case is defined as $\varepsilon = p_1 - p_0$. A joint likelihood function for ε and p_0 can be derived by substituting $p_0+\varepsilon$ for p_1 in Equation (6).

$$L(\varepsilon, p_0|s_0, f_0, s_1, f_1) = (p_0)^{s_0}\,(1-p_0)^{f_0}\,(p_0+\varepsilon)^{s_1}\,(1-p_0-\varepsilon)^{f_1} \qquad (7)$$

If it is reasonable to behave as though there were no prior knowledge linking p_0 and ε, a marginal likelihood function for ε can be obtained by integrating the function $L(\varepsilon, p_0)$ with respect to a marginal noninformative prior distribution for p_0. Because p_1 is a probability, then $0 < p_0 + \varepsilon < 1$, and the assumption of independence is an approximation; it is a very close approximation, however, for a wide range of possible values of p_0 and ε. The reasonableness of the assumption is only threatened when p_0 and approach 0 or 1, and the sample sizes are small. Furthermore, other measures of effect (e.g., change in odds ratio, percent change in rate, and relative risk) can be used to help achieve independence between p_0 and ε (Wolpert and Eddy 1986). If $g(p_0)$ is the (possibly noninformative) prior for p_0, the marginal likelihood function for ε is

$$L(\varepsilon|s_0, f_0, s_1, f_1) = \int (p_0)^{s_0} (1-p_0)^{f_0} (p_0+\varepsilon)^{s_1} (1-p_0-\varepsilon)^{f_1} g(p_0)\, dp_0 \tag{8}$$

There are several possible choices for noninformative priors for the parameter p_0. Use of a uniform prior and the normal approximation for the binomial likelihood function leads to the approximation

$$L(\varepsilon|s_0, f_0, s_1, f_1) = N(\mu, \sigma^2) \tag{9}$$

where $N(\mu, \sigma^2)$ is the normal distribution with mean $\mu = s_1/n_1 - s_0/n_0$ and variance $\sigma^2 = s_0 f_0/(n_0)^3 + s_1 f_1/(n_1)^3$. It can be shown that the parameters μ and σ^2 differ only by terms of order $(1/n_0 + 1/n_1)$ from those that would be obtained with other reasonable candidates for noninformative priors.

Equation (9) can be illustrated with the TIMI study, the results of which are given in Table I (TIMI 1985). Figure 1 shows the likelihood function for the effect on reperfusion of t-PA compared with IV SK calculated from Equation (9) with $s_0 = 44$, $f_0 = 78$, $s_1 = 78$, $f_1 = 40$. The horizontal axis of this figure is ε, the true effect of the technology (compared with IV SK) on reperfusion. There is no vertical scale because a likelihood function is defined only up to an arbitrary constant. The likelihood function can be multiplied by a noninformative prior and normalized to derive a posterior distribution that has a virtually identical appearance. It would show that t-PA (compared with IV SK) can be expected to increase the probability of reperfusion by about 30%, with a 95% range of confidence[9] from 18% to 42%.

Adjustment of Likelihood Functions for Biases. Adjustment of evidence for biases that affect their internal validity, or for factors that affect their applicability to a particular assessment (external validity) will be illustrated with formulas that adjust likelihood functions for RCTs to correct for errors in outcome measurement, crossover of patients, and two types of confounding factors. Formulas have also been written to adjust studies for differences in length of followup.

Adjustment for Errors in Measurement of Outcomes. In an RCT, the joint likelihood function for p_0 and p_1 is a function of the observed number of successes and failures in the treated and control groups. Unfortunately, there can be errors in the method used to determine whether an outcome is a success or failure, due to such problems as an imperfect measurement instrument, errors in reporting, clerical errors, and so forth. If the true outcomes were known, the counts of successes and failures could simply be corrected, and used in Equation (8). More frequently, it is only possible to estimate the chance of error. Let a_0 be the probability that, in the control group, a true success will be incorrectly labeled a failure, and let b_0 be the probability that, in the control group, a true failure will be labeled a success. Let a_1 and b_1 be the corresponding probabilities for the treated group. It is easy to show that

$$L(p_0, p_1|s_0, f_0, s_1, f_1, a_0, b_0, a_1, b_1) = \tag{10}$$

$$\left[p_0(1-a_0) + (1-p_0)b_0\right]^{s_0} \cdot \left[p_0 a_0 + (1-p_0)(1-b_0)\right]^{f_0} \cdot \left[p_1(1-a_1) + (1-p_1)b_1\right]^{s_1} \cdot \left[p_1 a_1 + (1-p_1)(1-b_1)\right]^{f_1}$$

Each of the expressions in brackets represents the true probability of a success or failure, as a function of the values of p_i, a_i and b_i.

The joint likelihood function for ε and p_0 $(L(\varepsilon, p_0))$[10] can again be obtained by substituting $\varepsilon + p_0$ for p_1 in Equation (10), and a marginal likelihood function for ε can be obtained by integrating $L(\varepsilon, p_0)$ with respect to a (possibly noninformative) prior distribution for p_0.

Any uncertainty about a_0, b_0, a_1, and/or b_1 can be described with probability distributions, and the expression integrated with respect to the uncertain variable(s), causing any uncertainty about any of these parameters to be encoded in the likelihood function. Thus, if uncertainty about a_0 is described by $h(a_0)$, then

$$L(p_0, p_1) = \int L(p_0, p_1|a_0)\, h(a_0)\, da_0 \tag{11}$$

Incorporating uncertainty about all four parameters (a_0, a_1, b_0, b_1) would involve quadruple integration.

This feature of the Confidence Profile Method can be illustrated by deriving a likelihood function for the reperfusion rates observed in the TIMI study, adjusting for the possibility that the observed reperfusion rates might have been distorted by the performance of the coronary angiography used to measure reperfusion.[11] Suppose we estimate $b_0 = 0.05$ and $b_1 - 0.05$ (based on estimates of the proportion of patients with occluded arteries that can be opened by angiography alone), and $a_0 = a_1 = 0$ (assuming that angiography will not close an open artery). Using these estimates of a_0, a_1, b_0, b_1, and Equation (10), the adjusted likelihood function for the effect of t-PA versus IV SK on reperfusion, derived from the TIMI study, is shown in Figure 2.

Intensity and Additive Bias. Often the circumstances of a study do not precisely match the circumstances of interest, making it difficult to compare studies or to apply them directly (without adjustment) to a particular assessment. Examples are differences in the population, the technology, the providers, or other factors that can modify a technology's effect. These differences can affect the interpretation of a study in two basic ways, to create what will be called an intensity bias and/or an additive bias.

A bias is an intensity bias if it has a proportional effect on the effectiveness of the technology. Specifically, a factor is said to affect the intensity of a technology if, whatever the effect of the technology in the absence of the factor (ε), the presence or modification of the factor causes the effect (call this ε') to be $\varepsilon' = \tau\varepsilon$, where τ is the measure of the magnitude of the bias caused by the factor. Thus, $\tau = 1$ implies a factor causes no intensity bias. Examples of factors that affect the intensity of a technology are the dose of a drug, frequency of an examination, skill of a provider, type of equipment, or susceptibility of a patient to a treatment. The notion of intensity is described in statements such as "this technology has improved 20% since the study was completed," and "the effect of this technology in community hospitals will be only 80% that observed in a research setting." Notice that if a technology has no effect ($\varepsilon = 0$), presence or modification of the factor will leave the technology with no effect, and if the technology is harmful, increasing its intensity will indicate it is more harmful.

An additive bias shifts the probability or magnitude of an outcome in the treated and/or control groups by a constant amount. Specifically, a factor is said to cause an additive bias if, whatever the true effect of the technology in the circumstances of interest (ε), the bias causes the observed effect (ε') to be $\varepsilon' = \beta + + \varepsilon$, where β is the amount of the additive bias. Thus, $\beta = 0$ implies no additive bias, and β can be positive or negative. An example of a factor that causes additive bias is any difference between the treated and control groups of a controlled trial that can modify the probability of the health outcome, even in the absence of the technology (e.g., age of the patient, severity of the health problem).

If a particular study is thought to be affected by one or more factors that create an intensity or additive bias, with estimates of the intensity (τ) and additive biases (β), the likelihood function for p_0, p_1, and therefore for ε can be found by substituting $\beta+\tau\varepsilon$ for ε in Equation (8). For example, in a randomized controlled trial the adjusted joint likelihood function for ε and p_0 would be given by

$$L(\varepsilon, p_0|s_0, f_0, s_1, f_1, \beta, \tau) = (p_0)^{s_0} (1-p_0)^{f_0} (p_0+\beta+\tau\varepsilon)^{s_0} (1-p_0-\beta-\tau\varepsilon)^{f_1} \tag{12}$$

Uncertainty about β or τ can be described with probability distributions and the likelihood function integrated over those variables. If there are several independent factors affecting additive bias and/or intensity bias, β and τ can be vectors.

Additional issues must be considered when adjusting for a bias that is uncertain or variable, and that affects more than a single experiment. One example is that several experiments can be affected by the same bias. Another example is that a factor can modify the effect of a technology on a health outcome, but the effect must be estimated from indirect evidence. In such a case it is not possible to adjust the bias simply by adjusting the likelihood function for an individual experiment (as in Eq. [12]).

In the first example the appropriate approach depends on whether the biases affecting the separate experiments are independent. Suppose there are n studies affected by the same biases β and τ, which have distributions $g(\beta)$ and $h(\tau)$. If the biases are independent, the likelihood function for combined evidence in the n studies is given by

$$L(p_0, p_1|X_{\cdot 1}, X_{\cdot 2}, \ldots X_{\cdot n}, \beta, \tau) = \tag{13}$$

$$\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L_1 (p_0, p_1|X_{\cdot 1}, \beta, \tau)\, g(\beta)\, h(\tau)\, d\beta\, d\tau\right] \cdots$$

$$\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L_n(p_0, p_1|X_{\cdot n}, \beta, \tau)\, g(\beta)\, h(\tau)\, d\beta\, d\tau\right]$$

If the biases are completely dependent, such that their magnitudes are the same in all n experiments, then

$$L(p_0, p_1|X_{\cdot 1}, X_{\cdot 2}, \ldots X_{\cdot n}, \beta, \tau) = \tag{14}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[L_1(p_0, p_1|X_{\cdot 1}, \beta, \tau) \cdots L_n(p_0, p_1|X_{\cdot n}, \beta, \tau)\right] g(\beta)\, h(\tau)\, d(\beta)\, d(\tau)$$

The second example can arise if all the evidence is gathered in experimental circumstances that are different from the circumstances of interest. In such cases, the approach is first to analyze all the evidence to derive a posterior distribution for the technology's effect in the experimental setting, call this $\pi(\varepsilon'|X_{\cdot\cdot})$, and then derive a posterior distribution for the technology's effect in the circumstances of interest by substituting $\varepsilon = \beta + \tau\varepsilon'$, for ε'. If there is uncertainty about β or τ, it can be incorporated by convolution. That is,

$$\pi(\varepsilon|X_{\cdot\cdot}) = g(\beta) * [h(\tau) \bullet \pi(\varepsilon'|X_{\cdot\cdot})] \tag{15}$$

where $*$ is the convolution operator and $\bullet$ is the multiplication operator for two distributions.[12] Use of Equation (15) will be illustrated below.

Dilution and Contamination. A frequent problem in controlled trials is that some subjects in the "treated group" might not receive the designated technology, which "dilutes" the observed effectiveness of the technology, and some subjects in the control group might inappropriately receive the technology, thereby "contaminating" the control group and distorting the observed effectiveness of the technology. In some cases, the number of subjects in each group who "cross over" and the number of successes and failures in each group are known. If it is reasonable to assume they are similar to the other subjects with respect to factors that can affect outcomes, the counts could be corrected and an appropriate likelihood function could be derived along the lines of Equation (8).

Let q_0 be the number of subjects in the control group who cross over to receive the technology, and let q_1 be the number of subjects in the group offered the technology (the "treated" group) who cross over and do not receive the technology. Further, let j and k be the number of successes in the crossover control group and the crossover treated group, respectively. Then

$$L(p_0, p_1) = \tag{16}$$

$$(p_0)^{s_0 - j + k} (1 - p_0)^{n_0 - s_0 - q_0 + j + q_1 - k} (p_1)^{s_1 - k + j} (1 - p_1)^{n_1 - s_1 - q_1 + k + q_0 - j}$$

As before, the marginal likelihood function for ε can be calculated by substituting $\varepsilon + p_0$ for p_1 to derive a joint likelihood function for ε and p_0, and then integrating that with respect to a (possibly noninformative) prior distribution for p_0.

A more complicated formula is required if there is reason to believe that subjects who cross over are *not* similar to the other subjects with respect to the expected effect of the technology. Let β_0 and τ_0 be additive and intensity biases, respectively, that cause the subjects in the control group who cross over to respond differently to the technology than subjects chosen randomly from the "treated" group. Let p_0' designate the probability of a success in this subset of the control group that crosses over to receive treatment. Then in these subjects the effect of the technology is $\varepsilon' = p_0' - p_0 = \beta + \tau\varepsilon = \beta + \tau(p_1 - p_0)$, and the probability of all success is $p_0' = \beta + \tau(p_1 - p_0) + p_0$. Similarly, let β_1 and τ_1 be additive and intensity biases that cause subjects in the treated group who cross over to respond differently to the technology than subjects chosen randomly from the control group. Let p_1' be the probability of success in this subset of the treated group. Then in these subjects the effect of the technology is $\varepsilon_1' = p_1 - p_1' = \beta_1 + \tau_1\varepsilon = \beta_1 + \tau_1(p_1 - p_0)$ and the probability of success is $p_1' = p_1 - \beta_1 + \tau_1(p_1 - p_0)$. Using this notation

$$L(p_0, p_1) = (p_0)^{a_0-j}(1-p_0)^{n_0-a_0-e_0+j}(p_0')^{j}(1-p_0')^{e_0-j} \tag{17}$$
$$(p_1')^{k}(1-p_1')^{e_1-k}(p_1)^{a_1-k}(1-p_1)^{n_1-a_1-e_1+k}$$

An expression in terms of β_0, β_1, τ_0 and τ_1 is easily obtained by substituting $\beta_0 + \tau_0(p_1 - p_0) + p_0$ for p_0' and $p_1 - \beta_1 + \tau_1(p_1 - p_0)$ for p_1' in Equation (17). As before, a likelihood function for ε can be obtained by substituting $p_0 + \varepsilon$ for p_1 and integrating over a (possibly noninformation) prior distribution for p_0. Uncertainty about the βs or τs can be described with probability distributions and the likelihood function integrated over those variables.

Equations (16) and (17) will not be illustrated here.

Multiple-Link Chains. Derivation of a Confidence Profile from indirect evidence involving multiple-link chains is performed in two steps: first evaluate evidence for each link in the chain to derive a probability distribution for that link, and then calculate across the links to obtain a probability distribution for the entire chain. Formulas for the first step have been described. Formulas for the second step will be given for a two-link chain involving dichotomous intermediate outcomes and health outcomes, taking into account the possibility that the intermediate outcome is not a perfect indicator for the health outcome.

Derivation of the formula is illustrated in Figure 3. Let T_0 represent the event the technology is not done (the control group); T_1 represent the event the technology is done (the treated group); I_0 represent the event the intermediate outcome does not occur; I_1 represent the event the intermediate outcome does occur; and H represent the event the health outcome occurs. Thus, the circles represent various events and combination of events (e.g., $[I_1, T_1]$ is the event that the technology is done $[T_1]$ and the intermediate outcome occurs $[I_1]$).

Define q as the probability of the intermediate outcome occurring, in the absence of the technology ($q = Prob(I_1|T_0)$), and p as the probability of the health outcome, in the absence of the intermediate outcome and the absence of the technology ($p = Prob(H|I_0, T_0)$). Let $\varepsilon_{t\omega}$ be the difference in the probability of the intermediate outcome (ω) with and without the technology (t) (the "effect" of the technology on the probability of the intermediate outcome), and let $\varepsilon_{\omega h}$ be the effect of the intermediate outcome (ω) on the health outcome (h), in the absence of the technology. That is, define

$$\varepsilon_{t\omega} = Prob(I_1|T_1) - Prob(I_1|T_0), \tag{18}$$

and

$$\varepsilon_{\omega h} = Prob(H|I_1, T_0) - Prob(H|I_0, T_0) \tag{19}$$

By these definitions

$$Prob(H|I_1, T_0) = p + \varepsilon_{\omega h} \tag{20}$$

Now let λ_{ij} be the difference in the probability of the j^{th} category of the intermediate outcome, caused by the presence of the technology (versus without the technology). That is

$$\lambda_{ij} = Prob(H|I_j, T_1) - Prob(H|I_j, T_0), \tag{21}$$

for j = 0 and j = 1.

This factor can be thought of as representing the inaccuracy of the intermediate outcome as a predictor of the health outcome. If the presence of the technology has no effect on the occurrence of the health outcome, other than to modify the probability of occurrence of the intermediate outcome, then the inaccuracy λ_{ij} will be 0.

Using this notation,

$$Prob(H|T_1) = \left[(q+\varepsilon_{t\omega})(p+\varepsilon_{\omega h}+\lambda_{i1})\right] + \left[(1-q-\varepsilon_{t\omega})(p+\lambda_{i0})\right], \tag{22}$$

and

$$Prob(H|T_0) = q(p + \varepsilon_{\omega h}) + p(1-q) \tag{23}$$

After simplifying,

$$\varepsilon = Prob(H|T_1) - P(H|T_0) = \varepsilon_{t\omega}\,\varepsilon_{\omega h} + \sum_j \lambda_{ij}\, Prob(I_j|T_1) \tag{24}$$

All of the parameters in Equation (24) can be distributions, either because they are estimated from empirical data and are therefore subject to sampling (and possibly subject to bias), or because they are estimated subjectively, with some uncertainty. In such cases the addition, subtraction, and multiplication operators in Equation (24) become the corresponding operators that apply to the distributions (e.g., addition of two scalars becomes convolution of two distributions). In symbols, let $\pi_{t\omega}(\bullet)$ be the distribution for $\varepsilon_{t\omega}$, $\pi_{\omega h}(\bullet)$ be the distribution for $\varepsilon_{\omega h}$, $f_{i0}(\bullet)$ be the distribution for λ_{i0} and $f_{i1}(\bullet)$ be the distribution for λ_{i1}. Let $\gamma_i = Prob(I_i|T_1)$, and $h_i(\bullet)$ distributions for γ_i, for $i = 0,1$. Then

$$\pi(\varepsilon) = \left[\pi_{t\omega}(\varepsilon_{t\omega}) \bullet \pi_{\omega h}(\varepsilon_{\omega h})\right] * \left[[f_{i0}(\lambda_{i0}) \bullet h_0(\gamma_0)] * [f_{i1}(\lambda_{i1}) \bullet h_1(\gamma_1)]\right] \tag{25}$$

where * is the convolution operator, and • is the multiplication operator.[13] Distributions for $f_{iu}(\lambda_{iu})$ and $h_i(\gamma_i)$ can be derived from empirical data by deriving a likelihood function for the parameter and multiplying by a noninformative prior to the parameter. When deriving distributions for the elements of Equation (25), care must be taken to ensure that they are independent—the same piece of evidence can not contribute to more than one distribution.

Example of Calculating Multiple-Link Chains. Equation (25) can be used to derive a probability distribution for the effect of t-PA compared with IV SK on one-year survival, using reperfusion as an intermediate outcome. A distribution for ε_{ti} is obtained from Equations (8) and (2), with s_0, f_0, s_1, f_1 estimated from the TIMI study ($s_0 = 44$, $f_0 = 78$, $s_1 = 78$, $f_1 = 40$) and a noninformative prior. Similarly, a distribution for ε_{ih} is obtained from Equations (8) and (2) using data reported by Kennedy (summarized in Fig. 4); $s_0 = 85$, $f_0 = 17$, $s_1 =14$, $f_1 = 0$, where the subscript 0 refers to patients who do not reperfuse, and the subscript 1 refers to patients who reperfuse.

Distributions for the "inaccuracy" of the intermediate outcome can be estimated from data that relate the intermediate outcome to the health outcome, with and without the thrombolysis (as defined in Eq. [18]). Kennedy's data (Fig. 4) again can be used to derive the needed distributions for λ_{t0} and λ_{t1}. The distribution is obtained by deriving a likelihood function for the difference in rates (see Eq. [8]) and multiplying by a noninformative prior (see Eq. [2]).[14]

Distributions for $Prob(I_0|T_1)$ and $Prob(I_1|T_1)$ are obtained in a similar fashion. First derive a likelihood function for the likelihood of the true rates of the event if interest (e.g., I_0, the nonoccurrence of the intermediate outcome), as a function of the observed rates. For example, if we denote $Prob(I_0|T_1)$ as γ_0, and $Prob(I_1|T_1)$ as γ_1, then

$$L(\gamma_i|s_i f_i) = (\gamma_i)^{s_i} (1-\gamma_i)^{f_i} \tag{26}$$

where s_i is the observed number of occurrences of the event of interest, and f_i is the number of nonoccurrences of the event. Then a posterior distribution for γ_i can be obtained by

$$h_i(\gamma_i|s_i f_i) = k\, L(\gamma_i|s_i f_i)\, \pi_{\gamma_i}(\gamma_i) \tag{27}$$

where $\pi_{\gamma_i}(\gamma_i)$ is a noninformative prior distribution for γ_i and k is a normalizing constant.[15]

With the necessary distributions for ε_{ti}, ε_{ih}, λ_{t0}, λ_{t1}, γ_0, and γ_1 estimated from data reported by TIMI and Kennedy, Equation (25) can be used to derive a probability distribution for the effect of t-PA (versus IV SK) on one-year survival, using reperfusion as an intermediate outcome. The result is shown in Figure 5.

Estimating Long-Term Outcomes from Short-Term Outcomes. The formulas for chains can be used to estimate long-term outcomes from data on short-term outcomes. This can be accomplished by constructing a chain that relates the technology to the short-term outcome (link 1), and the short-term outcome to the long-term outcome (link 2). The chain can be executed if there are data from previous research relating the short-term outcome to the long-term outcome.

This feature can be particularly useful in tracking the evolution of a technology. A long-term experiment can be conducted to evaluate the effect of a new technology, on both short-term and long-term outcomes. As variants of the technology are developed, short-term experiments can be conducted to test the effect of the new versions on short-term outcomes, while data from the original (long-term) experiment can be used for the second link.

Adjusting for Different Controls. Assessment of a health technology is often complicated by the fact that studies compare different variations of the technology with different controls. For example, Collen compared t-PA with conventional care, TIMI compared t-PA with IV SK, Kennedy compared IC SK with conventional care, and other RCTs have compared IV SK with conventional care, and IC SK with conventional care. In general, it is useful to think of a "family" of technologies, all intended to affect the same health outcomes (e.g., survival) for the same health problem (e.g., heart attacks). Each family would then consist of different variations of a basic type of technology (e.g., thrombolytic agents) and their controls, T_1, T_2,..., T_n. The existing evidence might compare any pair of technologies in the family.

Given whatever comparisons exist, the Confidence Profile Method can be used to derive Profiles for other comparisons that can be related by various independent pieces of evidence. This is performed by convolution. In general, let $\pi_{i,j}(\varepsilon_{i,j})$ be the Confidence Profile for the effect of Technology T_i compared with Technology T_j. If there is a Profile that relates Technology 1 to Technology 2, and another Profile that relates Technology 2 to Technology 3, then a Profile relating Technology 1 to Technology 3 can be derived by

$$\pi_{1,3}(\varepsilon_{1,3}) = \int \pi_{1,2}(\varepsilon_{1,2})\, \pi_{2,3}(\varepsilon_{1,3}-\varepsilon_{1,2})\, d\varepsilon_{1,2} \tag{28}$$

This use of the Confidence Profile Method will be illustrated below.

Comparing Different Technologies. A closely related problem is that existing evidence might relate two technologies in a family to a common third technology (e.g., to the same control), and a policymaker wants to compare the first two technologies *to each other*. For example we might have a Profile that related Technology 1 to 3 and another Profile derived from independent evidence that related 2 to 3, and want to derive a Profile that related Technology 1 to 2. This is also accomplished by convolution.

$$\pi_{1,2}(\varepsilon_{1,2}) = \int \pi_{1,3}(\varepsilon_{1,}) \, \pi_{2,3}(\varepsilon_{1,3} - \varepsilon_{1,2})\, d\varepsilon_{1,3} \tag{29}$$

Equation (29) will be illustrated below (Fig. 10).

Combining Chains. The evidence for a technology often involves more than one chain. For example, there might be direct evidence from one or more RCTs relating the technology directly to the health outcome (chain 1), as well as indirect evidence that the technology changes an intermediate outcome, which is related to the health outcome (chain 2). A probability distribution for the effect of a technology that combines both bodies of evidence is obtained by first deriving probability distributions for each of the chains separately, and then multiplying the two distributions, point by point (with an additional term that depends on the prior distribution for the technology's effect). Specifically let $\pi_1(\varepsilon|X_{\cdot 1})$ be the distribution for the technology's effect derived from the first chain, $\pi_2(\varepsilon|X_{\cdot 2})$ be the distribution derived from the second chain, and so forth for an arbitrary number of chains (n). We seek the distribution for $\pi(\varepsilon|X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot n})$, based on all n chains.

For each chain, by Bayes formula

$$\pi_i(\varepsilon|X_{\cdot i}) = k_i\, L_i(\varepsilon|X_{\cdot i})\, \pi(\varepsilon) \tag{30}$$

and therefore

$$L_i(\varepsilon|X_{\cdot i}) = 1/k_i\, \pi(\varepsilon|X_{\cdot i})\, 1/\pi(\varepsilon) \tag{31}$$

where $\pi(\varepsilon)$ is the (noninformative) prior distribution for ε, and k_i is a normalizing constant.

Furthermore,

$$\begin{aligned}\pi(\varepsilon|X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot n}) &= k\, L(\varepsilon|X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot n})\, \pi(\varepsilon) \\ &= k\, L_1(\varepsilon|X_{\cdot 1})\, L_2(\varepsilon|X_{\cdot 2}) \cdots L_n(\varepsilon|X_{\cdot n})\, \pi(\varepsilon)\end{aligned} \tag{32}$$

Substituting Equation (31) into (32)

$$\begin{aligned}\pi(\varepsilon|X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot n}) &= k\prod_i \left[1/k_i\, \pi_i(\varepsilon|X_{\cdot i})\, 1/\pi(\varepsilon)\right] \pi(\varepsilon) \\ &= k'\prod_i \left[\pi_i(\varepsilon|X_{\cdot i})\right]\left[1/\pi(\varepsilon)\right]^{n-1}\end{aligned} \tag{33}$$

where k and k' are normalizing constants.

ASSESSMENT OF THE EFFECT OF t-PA ON ONE-YEAR SURVIVAL

The evidence summarized in Table 1 and the formulas just given can be used to derive a Confidence Profile for the effect on one-year survival of t-PA versus conventional care (Eddy 1986). The result is shown in Figure 6, marked $\pi(\varepsilon|X_{\cdot\cdot})$.

Notice that none of the existing controlled trials examines this question directly; the Profile must be derived from the indirect evidence in Table 1, using the methods just described. The specific steps are (1) construct a two-link chain that relates t-PA (versus IV SK) to reperfusion, and reperfusion to one-year survival; (2) use the results of the TIMI study and Equation (8) to derive a probability distribution for the first link, the effect of t-PA (versus IV SK) on reperfusion; (3) use the results of the Kennedy study to derive a probability distribution for the second link of the chain (Eq. [19]), and to estimate the inaccuracy of reperfusion as a predictor of one-year survival (Eq. [21]);(4) combine the evidence about the links (and the connection between links) to derive a probability distribution for the effect of t-PA versus IV SK on a one-year survival (Eq. [25]); (5) combine the results of 20 other studies (summarized in Yusuf et al 1985) to derive a probability distribution that relates IV SK to conventional care (Eq. [8]) (see Eddy 1986 for details); (6) use Equation (28) to derive a probability distribution for the effect on one-year survival of t-PA versus conventional care (call this π_1). This distribution is illustrated in Figure 6, marked $\pi(\varepsilon|X_{\cdot 1})$.

Then (7) construct a new chain that related t-PA versus conventional care (instead of IV SK) to reperfusion, and reperfusion to one-year survival; (8) use the results of Collen's study to derive a probability distribution for the first link (Eq. [8]); (9) using probability distributions for the second link and the inaccuracy of reperfusion, calculate across the chain (with corrections for the inaccuracy of reperfusion) to derive a probability distribution for the effect of t-PA versus conventional care on one-year survival (Eq. .[25]) (marked $\pi_2(\varepsilon|X_{\cdot 2})$ in Fig. 6); and (11) combine π_1 and π_2 by Equation (33). The results is the Confidence Profile in Figure 6 marked $\pi(\varepsilon|X_{\cdot\cdot})$. This Profile combines the evidence in the TIMI, Collen and Kennedy studies, as well as 20 RCTs that compare IV SK with conventional care.

The Method can also be used to derive a Profile for the effect of IV SK (compared with conventional care). That Profile, the result of combining direct evidence from 20 RCTs in step 5 above, is shown beside the Profile for t-PA in Figure 7. (Because of the high degree of certainty about the effectiveness of IVSK, the scale for Fig. 7 is twice as high as the scale for the other figures.)

Adjustment for Intensity Bias. The use of the Confidence Profile Method to adjust for possible biases can be illustrated with a problem that arises in the interpretation of the evidence on t-PA. The effect of t-PA observed in published RCTs might underestimate the true effect of t-PA in realistic settings, because in the trials patients were catheterized before administration of t-PA (to observe perfusion), which delayed administration of the drug, which in turn might have decreased its effectiveness. The impact of this possibility can be included in the assessment by estimating how much more effective *t-PA* might be in actual clinical settings. (This estimate can be based on animal studies, knowledge of clotting mechanisms, knowledge of the mechanism of action of the drug, and review of human studies that recorded outcomes as a function of time.) For example, if the trials are believed to understate the true effect of t-PA by about 20% (implying an intensity bias in the trials of $\tau = 0.8$), the Profile marked $\tau = 0.8$ in Figure 8 would be obtained by applying Equation (12). If there were uncertainty about the estimated intensity bias described, say, by a beta distribution with a mean = 0.8 and variance = 0.2, the Profile marked $\tau = 0.8u$ in Figure 9 would be obtained. Figure 9 also includes for comparison the original, unadjusted Profile (marked $\tau = 1$).

Comparing Generations of Technologies. Use of the Confidence Profile Method to compare different variations of a technology is illustrated in Figures 6 and 10. Figure 6 showed the effects of two individual versions (t-PA and IV SK), both compared with conventional care. Figure 10 shows the effect of t-PA compared with IV SK, calculated from Equation (29).[16]

RESEARCH PLANNING

Once Confidence Profiles for the effects of a technology on various outcomes have been derived, they can be used for a variety of purposes, such as adjustment for risk aversion, comparison of a technology's benefits and harms, derivation of a measure of overall benefit, and research planning. This section uses the results of the assessment of t-PA compared with conventional care to illustrate one of these uses—research planning.

Research planning is a complicated activity, ideally involving estimating the probability a new experiment will yield particular results, the probabilities those particular results will change behavior, and the change in health outcomes expected from the change in behavior. The Confidence Profile itself provides an estimate of the third element—how use of the technology is expected to change health outcomes. The Confidence Profile Method can also estimate the first element—the probability a particular experiment will yield a result that will change behavior. For convenience we will call this a "Delta Result" or Δ Result (drawing on the common use of the Greek letter "Δ" to denote a difference or change). Notice that a single experiment can produce several different Δ Results, depending on the type of action the result triggers, and the force with which it triggers it (i.e., the proportion of people who will change behavior if the result occurs). For example, if an experiment indicates a technology causes a 60% increase in survival, 99% of physicians might adopt it, whereas if the experiment indicates a 5% increase in survival, only 10% of physicians might adopt it. When the chance of a Δ Result and the Profile for the effect of a change in use of the technology are combined with estimates of the second element—how a Δ Result will change use of the technology—the impact of the experiment on health outcomes can be calculated, different experiments compared, and priorities set.

Use of the Confidence Profile Method to calculate the probability a particular experiment will yield a Δ Result is illustrated for an RCT, using a particular Δ Result—a statistically significant result. The principles behind the calculations are as follows. Let δ be the observed difference between the rate in the treated and control groups of an RCT (thus $\delta = s_1/n_1 - s_0/n_0$), let ε be the true difference, and let $f(\delta|\varepsilon)$ be the distribution for the observed difference, given a true difference of ε. Define μ to be the threshold for determining statistical significance of an RCT under the null hypothesis of no effect.

μ will depend on the level of statistical significance chosen, and whether a one-tail or two-tail test is being performed. For example, if a level of significance of $p = 0.05$ is chosen, and if we are performing a one-tail test for a positive difference in rates of a dichotomous outcome, then μ is found by solving the following equation for μ

$$0.05 = \int_{\mu}^{1} f(\delta|\varepsilon = 0)\, d\delta \qquad (34)$$

To calculate the probability that an RCT of particular size n_0 and n_1 will yield a statistically significant result, first derive a distribution for the outcome of the trial, based on the current distribution for the true effect of the technology (the Confidence Profile $\pi(\varepsilon|X_{..})$). This distribution is

$$\int_{-1}^{1} f(\delta|\varepsilon)\, \pi(\varepsilon|X_{..})\, d(\varepsilon) \qquad (35)$$

For the one-tail test just described, the probability a trial will yield a statistically significant result is

$$\int_{\mu}^{1} \int_{-1}^{1} f(\delta|\varepsilon)\, \pi(\varepsilon|X_{..})\, d\varepsilon\, d\delta \qquad (36)$$

To apply Equation (36) we need a distribution for $f(\delta|\varepsilon)$. This will depend on the experimental design and the available empirical data. For example, if the contemplated trial is an RCT, it is appropriate to expand $f(\delta|\varepsilon)$ over p_0 to obtain

$$f(\delta|\varepsilon) = \int_{0}^{1} f(\delta|p_0\, \varepsilon)\, g(p_0)\, dp_0 \qquad (37)$$

where $g(p_0)$ is a prior distribution for p_0. For sufficiently large sample sizes, $f(\delta|p_0\, \varepsilon)$ is well approximated by a normal distribution with mean $\mu = \varepsilon$ and variance $\sigma^2 = [p_0(1 - p_0)]/n_0 + [(p_0 + \varepsilon)(1 - p_0 - \varepsilon)]/n_1$. where n_0 and n_1 are the number of observations in the control and treated groups of the contemplated experiment. A distribution for p_0 can be derived from the currently existing evidence using the methods described in previous sections (e.g., Eqs. [26] and [27]).

Application of Equation (36) is illustrated with an analysis of the probability that RCTs of various sizes will yield statistically significant results (one-tail, $p = 0.05$), using the Profile for the effect of the technology illustrated in Figure 6. The results are shown in Figure 11. For example, given the existing evidence about the effect of t-PA versus conventional care, the probability a new RCT with a total of 1200 patients would show a statistically significant increase in one-year survival is about 80%. If the trial were to simulate "actual practice" and not involve catheterization, the probability of a statistically significant result would be higher because the Profile for the technology in this circumstance shows a greater effect (see Figs. 8 and 9). This will not be illustrated here.

This section has focused on calculating the probability of a statistically significant result. Recall that the method is more general, and can be applied to a wide variety of Δ Results. Examples of other Δ Results are that the experiment will show the technology has a "positive" effect ($\delta > 0$), the experiment will show an effect between, say, 0 and 10% ($0 < \delta \leq 0.10$), the experiment will show an effect greater than 10% (($\delta > 0.10$), and so forth.

DISCUSSION

The health of millions of people and the expenditure of billions of dollars depend on the decisions of health care practitioners and policymakers about the appropriate use of medical technologies. If these decisions are not to be arbitrary, they should be based on estimates of the effects or outcomes of the technologies—what good or harm they can be expected to cause. Making these estimates accurately, however, can be extremely difficult. The traditional approach involves collecting individual pieces of evidence, and "synthesizing" their results into a conclusion by a single global subjective judgment. The result of this process is usually a statement such as "the technology should be used for the following indications..." Rarely is there an explicit description of the expected magnitude of the technology's effect, much less a description of how that magnitude was estimated or the range of uncertainty.

For complicated assessment problems that involve many pieces of evidence, evidence from studies with different designs, indirect evidence involving intermediate outcomes, biases, or other complicating factors, this approach is vulnerable to oversimplification, errors in reasoning, and wishful thinking.

The Confidence Profile Method was developed to provide a formal framework and formulas for adjusting and combining evidence, and incorporating subjective judgments, to estimate a technology's effect on outcomes. The Method breaks the process of evaluating evidence into parts—down to the level of individual chains, individual pieces of evidence, and individual biases—and then combines the parts. The result is a quantitative (and visual) description of a technology's effect, both the magnitude of the effect and the range of uncertainty about the effect.

Depending on the available evidence, other techniques are available for making quantitative estimates of the technology on a health outcome. If the evidence consists of a single RCT that compares the designated technology with the designated control, in circumstances that match the circumstances of interest, standard statistical methods can be used to estimate the effect of the technology and confidence limits.[17] If the evidence consists of several RCTs that all compare the same technology with the same control in the circumstances of interest, their results pooled, again yielding an estimate of the effect and confidence limits. If there are several RCTs, but some differ with respect to the recipients or other confounding factors, these differences can sometimes be adjusted for by stratification and related statistical techniques (Kleinbaum et al 1984, Anderson et al 1980). Meta-analysis[18] can be used to analyze a collection of RCTs and calculate an estimate and confidence limits for the "effect size" of a technology[19] (Glass 1977; Hedges and Olkin 1985). A collection of RCTs involving dichotomous outcomes can be analyzed to calculate a combined odds-ratio and confidence limits for the odds ratio (Mantel and Haenszel 1959; Mantel 1966; Peto et al 1977).

All these methods imply direct evidence from controlled trials. When this type of evidence exists, the Confidence Profile Method can also derive probability distributions for the effects of technologies, measured in a variety of ways such as the difference in probabilities, the odds ratio, the percent change in outcome rate. However, there is a large class of technologies for which the existing evidence is not suitable for analysis by other techniques, but that can be analyzed with the Confidence Profile Method. Some features of this class of technologies are: (1) there are no RCTs—the assessment must be based on one body of evidence relating the technology to intermediate outcomes and/or followup actions, and other evidence relates the intermediate outcomes (and followup actions) to health outcomes; (2) there are RCTs, but they differ from each other with respect to the technology being assessed, the control, the population, the providers, or other important features; (3) there are multiple studies with different designs (e.g., RCTs, nonrandomized controlled studies, clinical series, case-control studies, cross-sectional studies); and (4) interpretation of individual pieces of evidence is complicated by errors in outcome measurement, crossover of patients between "treated" and control groups, differences in length of followup, and other important factors. The assessment of t-PA illustrates many of these features.

The Confidence Profile Method also differs from other methods by its formal incorporation of subjective judgments. It is important to understand the role of subjective judgment in the Confidence Profile Method. Specifically, the Confidence Profile Method does not *require* subjective judgments. Rather it *enables* decisionmakers to incorporate subjective judgments should they feel a need to do so. If the evidence is clearcut and decisionmakers are content to accept it at face value, the Confidence Profile Method can be used to derive a posterior distribution for the technology's effect without using any subjective judgments (other than the initial judgment that the evidence is "clearcut," and other than the choice of a noninformative prior). If on the other hand, a decisionmaker identifies factors that influence the interpretation of the evidence, and if the decisionmaker wants to incorporate subjective judgments about these factors in the interpretation of the evidence, the Confidence Profile Method provides a formal language for accomplishing that.

It will never be possible to completely eliminate the need for subjective judgments in the evaluation of health technologies. The Confidence Profile Method attempts to improve the use of subjective judgment in several ways. First, by providing a formal framework for breaking the assessment problem into parts, the Method decreases the demands on subjective judgments. Instead of requiring policymakers and experts to make global judgments about dozens of factors all at once (e.g., "Should t-PA be used for patients with acute MI?"), the Method allows them to focus on one factor at a time (e.g., "By how much (what proportion) does a 60-minute delay in administration of t-PA reduce its effectiveness in increasing survival after an acute MI?"). Second, the judgments are targeted at elements that are intuitively accessible, and for which there is usually some supporting empirical evidence or practical experience. Third, the Method allows anyone who is uncertain about a parameter to express that uncertainty as a probability distribution. The uncertainty thus expressed about any parameter will be carried by the formulas through the entire analysis, automatically (according to the axioms of probability theory) combined with any uncertainty about any other parameters, and included in the final Confidence Profile. Fourth, the Method makes assumptions and judgments explicit, allowing for review. Fifth, it provides a formal language for combining subjective judgments; experts can think together about a parameter and describe their collective beliefs in a probability distribution. Disagreements can be explored by performing separate assessments, or resolved by describing a bimodal distribution for the parameter in question. Last, the Method can be used to estimate the value of additional information about a parameter that must be estimated subjectively.

Once derived, the Confidence Profiles have several uses in the design of health policies. The Profiles themselves provide explicit, visual descriptions of the effect of the technology—including the range of uncertainty about the effect—for use by decisionmakers (patients, practitioners, and policymakers). They can be revised to test the impact of different assumptions or beliefs about a variable, or to tailor an assessment to a particular set of circumstances (defined by a particular set of parameters). Because the Profiles provide a quantitative description of the uncertainty about a technology's effect, they enable the use of formal methods for incorporating risk aversion (e.g., calculating certainty equivalents and expected utilities). The Profiles (or certainty equivalents) provide a basis for comparing a technology's benefits and harms (using multidimensional utility theory), and enable the derivation of a quantitative measure of "overall" benefit (or harm). The quantitative measure of benefit also provides a basis for estimating a technology's marginal returns, and therefore for setting priorities. Profiles based on existing evidence can be used to estimate the value of conducting additional empirical research to collect more empirical evidence. Finally, a Profile can be continually revised to incorporate new evidence about a particular aspect of the assessment (e.g., new direct evidence, new indirect evidence, or new evidence about a particular parameter [e.g., bias] in the assessment).

The Confidence Profile Method (as do all methods of technology assessment) depends on the quality of the available evidence. While the Confidence Profile Method can combine evidence from many sources, and can adjust evidence for a wide variety of biases, it can not create evidence where it does not exist. Like all methods of technology assessment, the value of the Confidence Profile Method is improved if the volume and quality of empirical research is improved.

FOOTNOTES

[1] The method itself is more general, being applicable to the assessment of evidence about the effect of a wide variety of interventions on a wide variety of outcomes. However, the development of the Confidence Profile Method was initially stimulated by a need to assess health technologies.

[2] Some outcomes can be both health outcomes (people care about them) and intermediate outcomes (they are physiological variables that indicate the probability of other health outcomes.) An example is obesity.

[3] In general, for each link, the antecedent event (on the left) will be called an "action" and the subsequent event (on the right) will be called an "outcome." Thus for a two-link chain involving one intermediate outcome: for the first link the "action" is the performance of the technology, and the "outcome" is the intermediate outcome; for the second link the "action" is the occurrence of the intermediate outcome, and the "outcome" is the health outcome.

[4] This assumes the pieces of evidence are independent. If pieces of evidence are not independent, a joint likelihood function for the dependent pieces of evidence must be derived before use of Equation (1). Equation (1) also assumes that there is a particular true value of that all the pieces of evidence are trying to estimate. If there is reason to believe this is not the case—that the effect being estimated by one experiment is different than the effect being estimated by another—hierarchial Bayesian methods can be used to estimate a distribution for the true effects (Wolpert and Eddy 1986).

[5] For example, for dichotomous health outcomes and intermediate outcomes, this condition requires that $Prob(H|I,T) = Prob(H|I,T_0) = Prob(H|I)$, *where* H *is the occurrence of the health outcome,* I *is the occurrence of the intermediate outcome,* T *is the performance of the designated technology and* T_0 *is the designated control.*

[6] The two distributions should be independent in the sense that each is derived from different pieces of evidence.

[7] These judgments are called "focused" because they are made one at a time about specific elements of an assessment.

[8] As with standard statistical methods, sensitivity analyses might still be required to examine the importance of structural assumptions, such as the choice of a statistical model for a particular experiment.

[9] The 95% "range of confidence" is defined here as the range that has a 95% posterior probability of containing the true value. It is not the same as a confidence interval or confidence limit (see footnote 17).

[10] Where there is no danger of ambiguity, the parameters on which the likelihood function is conditioned will not be listed. Thus in this case $L(\epsilon, p_0) = L(\epsilon, p_0 | s_0, f_0, s_1, f_1, a_0, b_0, a_1, b_1)$

[11] Coronary angiography involves placing a catheter at the opening of the coronary artery and injecting a dye. Either the catheter or the dye could possibly open an occluded artery.

[12] Let X and Y be random variables with distributions $f_X(x)$ and $f_Y(y)$, respectively. The distribution for the random variable W = X+Y, denoted as $f_W(w)$ is calculated as

$$f_W(w) = f_X * f_Y = \int_{-\infty}^{\infty} f_X(x) f_Y(w-x) dx.$$

The distribution for the random variable $Z = XY$, denoted $f_Z(z)$, is calcuated as

$$f_Z(z) = f_X * f_Y = \int_{-\infty}^{\infty} \frac{1}{|x|} f_X(x) f_Y(z/x) dx$$

[13] See footnote 12.

[14] The distribution for λ_0 is calculated from Equation (8) using $s_0 = 35$, $f_0 = 6$, $s_1 = 85$, $f_1 = 17$. Similarly, λ_1 is calculated from Equation (8) using $s_0 = 88$, $f_0 = 5$, $s_1 = 14$, $f_1 = 0$.

[15] In this case the distribution for *Prob* $(I_0|T_1)$ is calculated from Equations (26) and (27) with $s_0 = 41$, $f_0 = 93$, and the distribution for $P(I_1|T_1)$ is calculated from Equations (26) and (27) with $s_1 = 93$ and $f_1 = 41$.

[16] The Profile in Figure 10 is different from the Profile in Figure 5 (indicating greater certainty and a slightly greater effect, because the former is based on the results of the TIMI study, while the latter incorporates information from both the TIMI study and Collen's study.

[17] "Confidence limits" (and "confidence intervals") do not define a probability distribution for a parameter. Confidence limits can be thought of as defining the set of null hypotheses that will cause the

observed results of a trial to be not statistically significant at a specified significance level.

[18] The term "meta-analysis" is often used in two senses. It is the name of a specific technique for combining evidence to estimate the effect size. It has also been used as a general term for the entire class of techniques used to combine evidence from many sources. Here the term is used in the restricted sense.

[19] The effect size is defined as the difference in magnitude or rate of the outcome with and without the technology, divided by the standard deviation of a parameter in the control group.

REFERENCES

Eddy, D.M. 1986. The Use of Confidence Profiles to Assess Tissue-Type Plasminogen Activator. Chapter in *Acute Coronary Care* 1987 G.S. Wagner and R. Califf (Eds). Martinus Nijhoff Publishing Company.

Eddy, D.M. and R. Wolpert. Extensions of The Confidence Profile Method for Technology Assessment I, Center for Health Policy Research and Education Working Paper, 1986 (in preparation).

Wolpert, R. and D.M. Eddy. Extensions of The Confidence Profile Method for Technology Assessment II, Center for Health Policy Research and Education Working Paper, 1986 (in preparation).

Jeffreys, H. *Theory of Probability* (3rd Edn.). Oxford University Press, London, 1961.

Bernardo, J.M. Reference Posterior Distributions for Bayesian Inference (with discussion). *J. Royal Statist. Soc.* 41 113–147, 1979.

Collen, D., E.J. Topol, A.J. Teifenbrunn et al. 1984. Coronary Thrombolysis with Recombinant Human Tissue-Type Plasminogen Activator: A Prospective, Randomized, Placebo-Controlled Trial. Circulation 70, 1012–1017.

Hedges, L.V., I. Olkin, 1985. *Statistical Methods for Meta-Analysis*. Academic Press, London.

Kennedy, J.W., J.L. Ritchie, K.B. Davis, et al. 1985. The Western Washington Randomized Trial of Intracoronary Streptokinase in Acute Myocardial Infarction. A 12-Month Follow-up Report. NEJM 312, 1073–1078.

Kleinbaum, D.G., L.L. Kupper, H. Morgenstern. 1984. *Epidemiologic Research, Principles and Quantitative Methods*. Lifetime Learning Publications, Belmont, CA. pp 343-351.

Anderson S., Auquier A., Hauck W.W. et al. *Statistical Methods for Comparative Studies. Techniques for Bias Reduction*. New York, John Wiley & Sons, 1980.

Mantel, N. 1966. Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration. Cancer Chemother Rep 50, 163.

Mantel, N. and W. Haenszel. 1959. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. JNCI 22, 719–748.

Peto, R., M. C. Pike, P. Armitage et al. 1977. Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient. *Br. J. Cancer* 35, 1–39.

Glass, G. V. 1977. Integrating Findings: The Meta-Analysis of Research, in L. Shulman (Ed) *Review of Research in Education* Vol. 5, Itasca, IL, Peacock.

TIMI Study Group. 1985. The Thrombolysis in Myocardial Infarction (TIMI) Trial. Phase I Findings. NEJM 312, 932–936.

Verstraete, M., R. Bernard, M. Bory et al. 1985a. Randomised Trial of Intravenous Recombinant Tissue-Type Plasminogen Activator, versus Intravenous Streptokinase in Acute Myocardial Infarction. *Lancet 1*, 842–847.

Verstraete, M., W. Bleifeld, R. W. Brower et al. 1985b. Double-Blind Randomised Trial of Intravenous Tissue-Type Plasminogen Activator versus Placebo in Acute Myocardial Infarction. *Lancet 2*, 965–969.

Yusuf, S., R. Collins, R. Peto et al. 1985. Intravenous and Intracoronary Fibrinolytic Therapy in Acute Myocardial Infarction: Overview of Results on Mortality, Reinfarction and Side-Effects from 33 Randomized Controlled Trials. *European Heart J.* 6, 556–585.

TABLE I: EVIDENCE FOR T-PA ANALYSIS

	Comparison			Rates			Statistical
Study	Treatment	Control	Outcome	Treatment	Control	p-value	Significance
TIMI	t-PA	IV SK	reperfusion*	78/118	44/122	0.001	yes
			in-hospital mortality†	7/143	12/147	0.1	no
Collen	t-PA	placebo	reperfusion	25/33	1/14	0.001	yes
Verstraete	t-PA	IV SK	perfusion	43/61	34/62	0 054	no
			in-hospital mortality	3/64	3/65	0 5	no
Verstraete	t-PA	placebo	perfusion	38/62	13/62	0.0001	yes
			in-hospital mortality	1/64	4/65	0 365	no
Kennedy	IC SK	CC‡	reperfusion	80/134	14/116	0.0001	yes
			12-month mortality	11/134	17/116	0 107	no

* Rates are given for patients with partial or total occlusion.

† Rates are given for all patients entered into the study, including patients with no occlusion.

‡ CC = conventional care

FIGURE 1

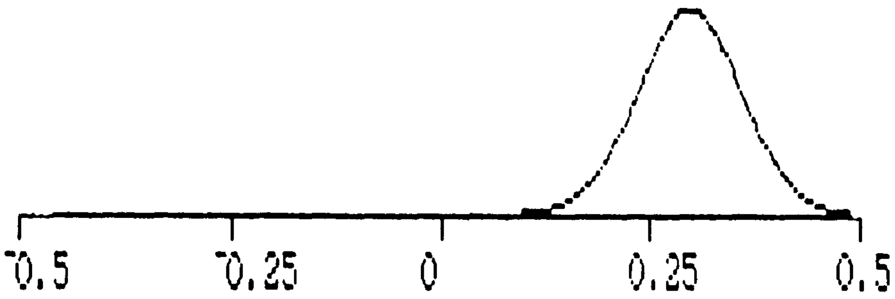


FIGURE 2

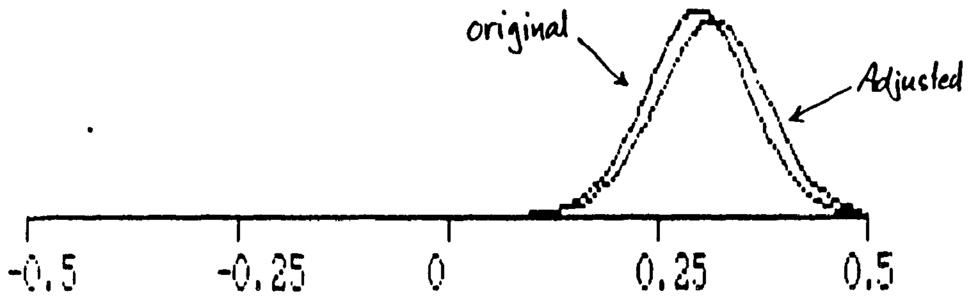


FIGURE 3

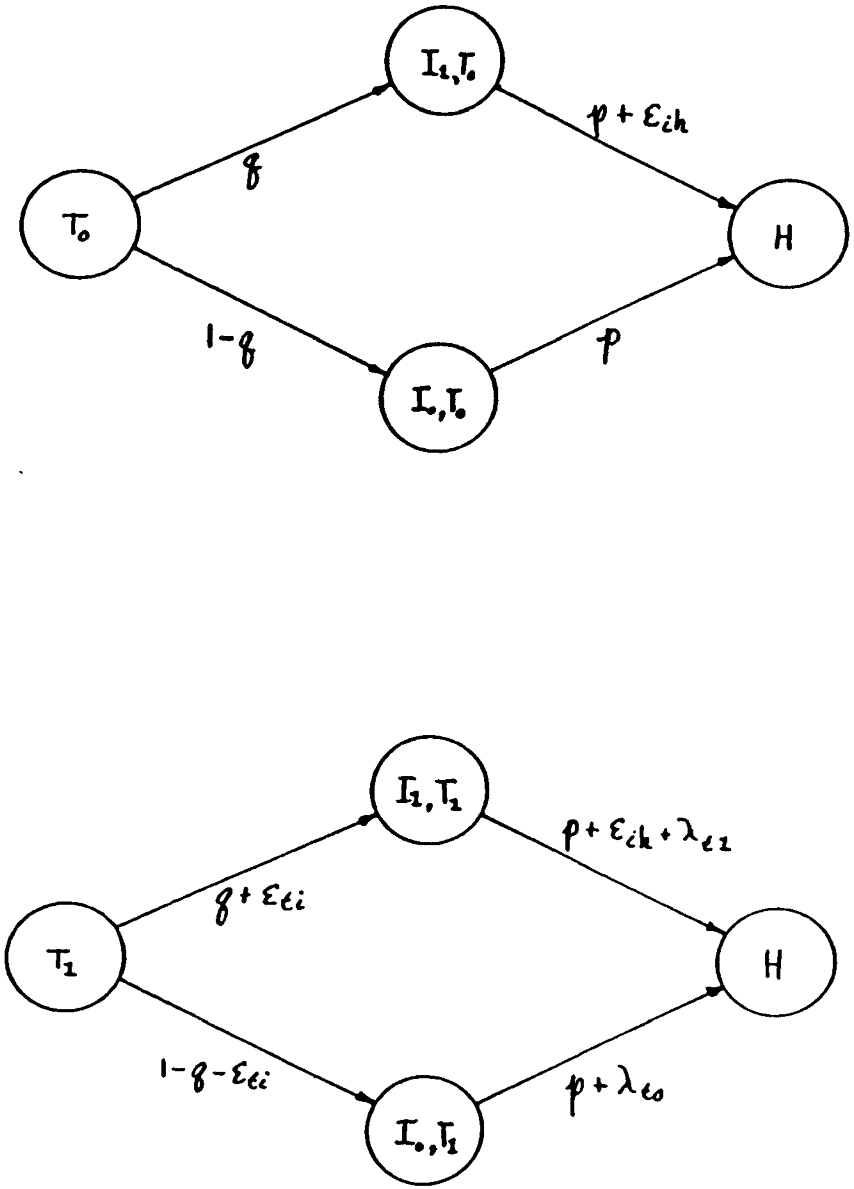


FIGURE 4

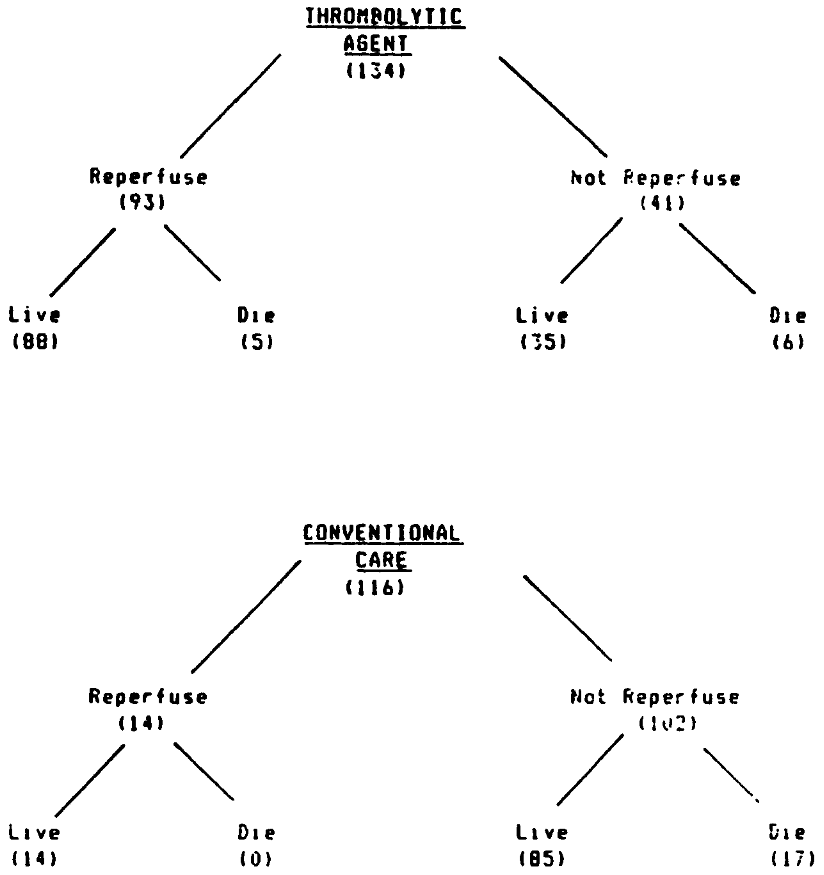


FIGURE 5

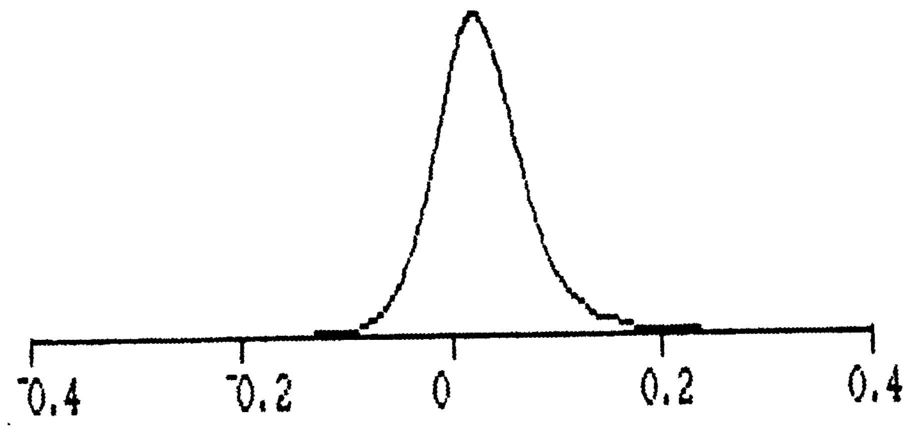


FIGURE 6

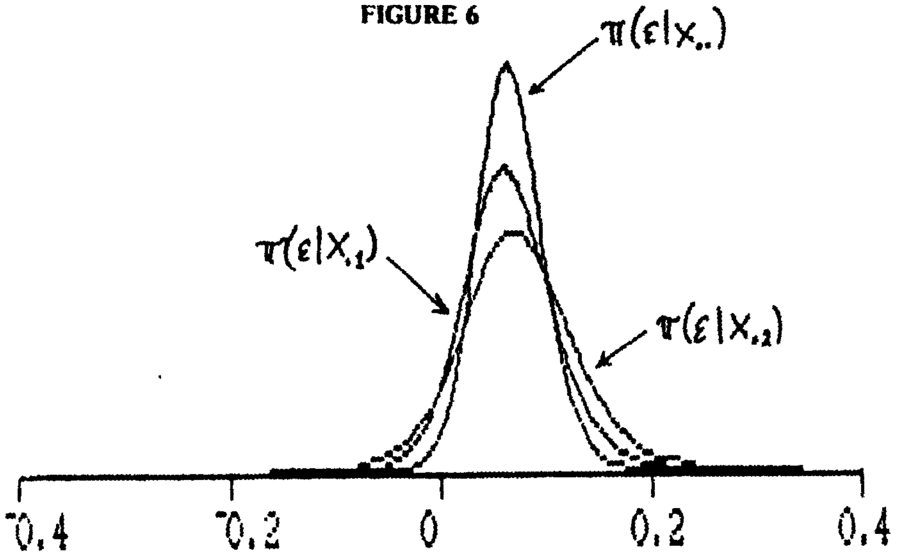


FIGURE 7

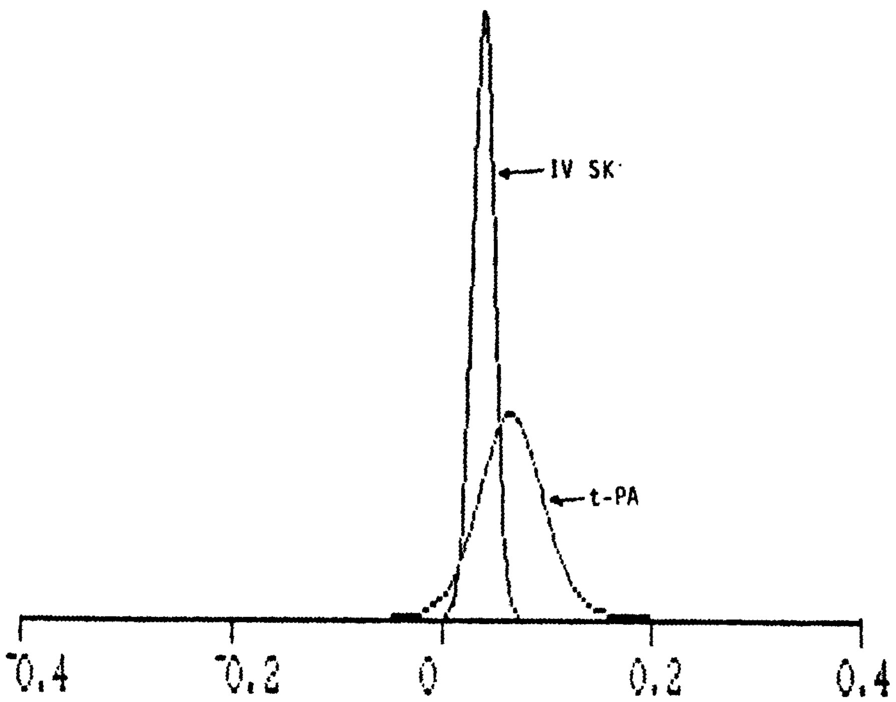


FIGURE 8

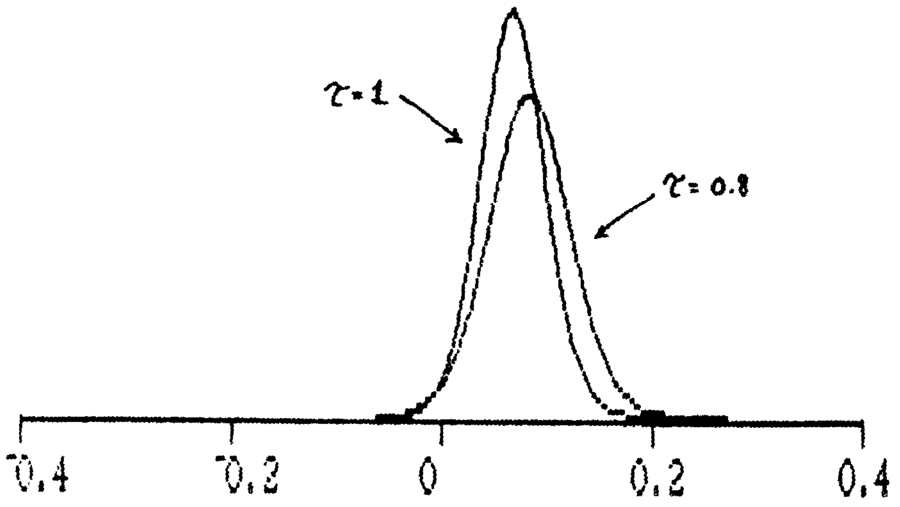


FIGURE 9

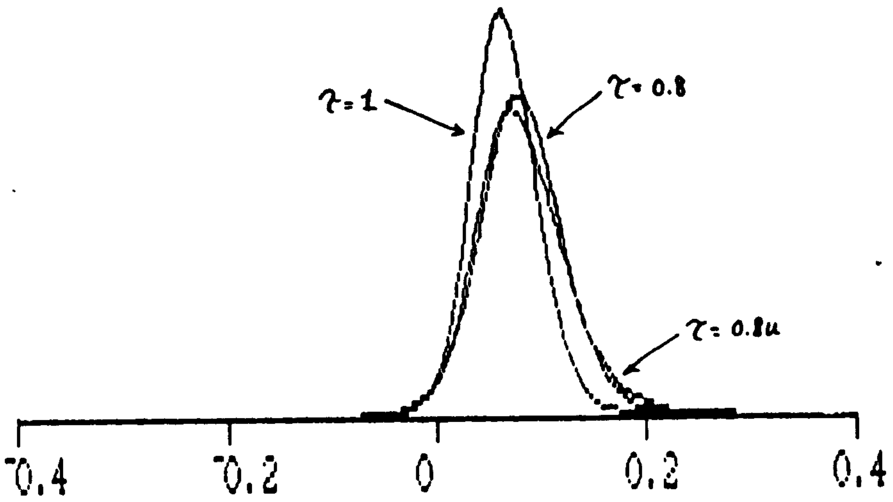


FIGURE 10

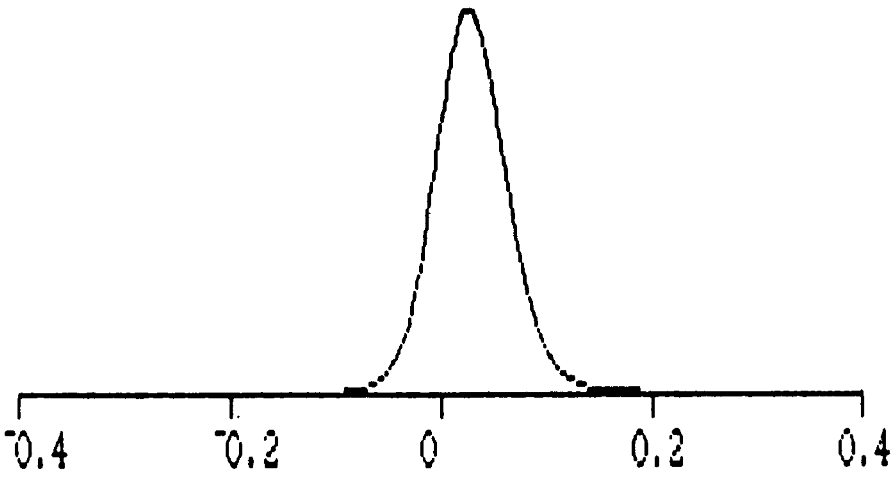
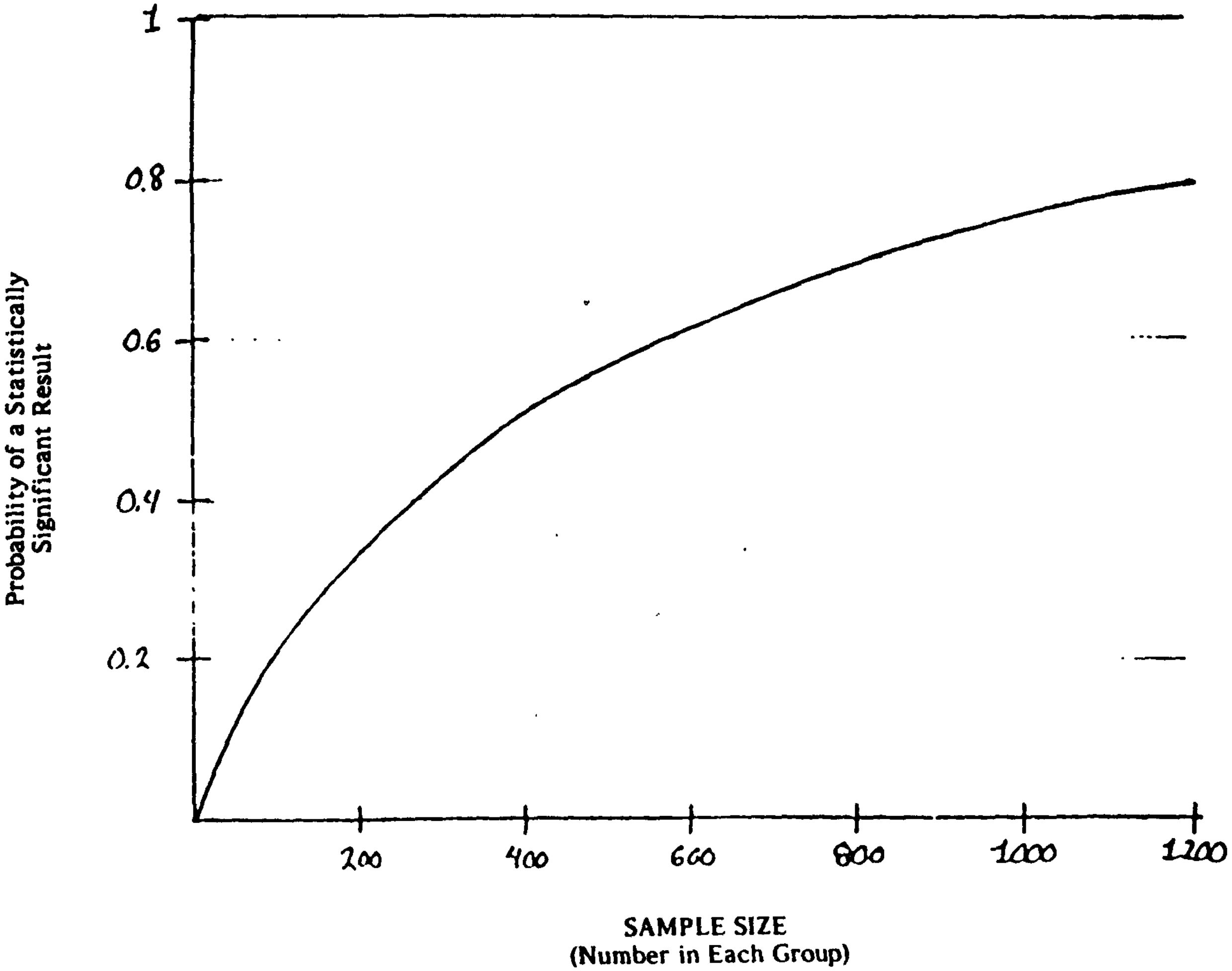


FIGURE 11


1
0.8
0.6
0.4
0.2
Probability of a Statistically Significant Result
200
400
600
800
1000
1200
SAMPLE SIZE
(Number in Each Group)

Choosing a Measure of Treatment Effect

Robert L Wolpert, Duke University

1. Introduction

Any new drug, surgical procedure, or other medical treatment must be shown to be effective before many insurance companys will reimburse for its use, and hence before it can become part of general medical practice. Before it can displace existing alternative treatments, it must be shown to have some advantage— to be more effective, less expensive, safer, or more convenient. Showing that such a treatment is effective at all, or more effective than existing treatments, requires evidence.

Evidence about the effectiveness of an experimental treatment can take many different forms, depending upon the design of the experiment intended to measure or detect treatment effect; the simplest evidence to analyze is that from a well-designed randomized controlled trial (RCT).

In such a trial subjects are randomly assigned to one of two or more groups. Usually one group (called the control group) receives "conventional care" (the standard and expected treatment at the time of the trial) while another (called the treated group) receives the experimental treatment, but in all other respects the patient protocols are identical for the two groups. In more complicated designs several groups might be given different treatments or different variations of a single treatment, all to be compared simultaneously.

The evidence from the trial consists of recorded measurements of experimental quantities for each subject. From this evidence the investigator can try to detect and quantify any systematic difference between the treated group and the control group; since the patient protocols were otherwise identical, such a difference must be attributed to either chance variation (from the random assignment of subjects to the groups) or to the treatment.

A systematic difference between the groups could be caused by improvement (or harm) caused by the treatment, by chance variations in the study populations, by side effects of the treatment or control protocols, or even by differing sample sizes in the treated and control groups. The investigator must choose a measure of treatment effect which is sensitive to the expected improvement or harm caused by the treatment, and relatively insensitive to unimportant side-effects and to chance variations.

The large sample sizes necessary to minimize chance fluctuation are not always attainable in medical trials. This forces us to pay careful attention to the probability distribution of the chance variations due to random sampling and may lead us to consider combining the evidence from multiple trials. When the evidence from a single trial is inconclusive, or when the evidence from several sources seems to be contradictory, we might like to pool the evidence from more than one trial and make inferences on the basis of a synthesis of all available evidence.

The investigator has more freedom in the choice of an effect measure when inference about treatment effect is be made using the evidence from a single trial than when evidence from several studies must be combined. We will see below that it is sometimes possible for an investigator to choose a measure of effect which is comparable across studies, despite the inevitable variations in patient population and treatment detail, and to pool the evidence using objective Bayesian methods.

The problem of synthesizing evidence from multiple trials is a statistical minefield where at every step we are tempted to make assumptions and simplifications which can threaten the validity of our analysis. Yet without making some of these assumptions we can make no progress at all. We have to assume that the treatments studied in the several trials are in some way comparable, for example, and that the effect of treatment can be compared meaningfully despite differences among the experimental conditions (such as patient populations) of the trials. We usually make broad assumptions about the stochastic independence of evidence from separate trials, and also of the conditional independence of study results within different arms of each single trial. This note illustrates a simple fact— that the "technical" assumptions made to simplify a statistical analysis often have real consequences. Sometimes we must adapt our method of analysis to assure that the assumptions are not flagrantly false, in order to assure that our findings will accurately reflect what was observed.

In Section 2 the ideas are introduced in a simple example, the binomial RCT, in which only a single bit of data is taken for each subject. The formalities and notation necessary for the case of evidence from more general designs of clinical trials is considered elsewhere (Wolpert and Berger (1986), Wolpert and Eddy (1986)). A summary follows in Section 3. I would like to thank David Eddy for introducing me to the idea of using Bayesian methodology to combine evidence from clinical trials. The present work grew out of my efforts to understand and to extend the scope of his method of Confidence Profiles (see Eddy 1986).

2. The Binomial Trial

A binomial trial is an RCT in which the only experimental quantity measured for each subject is whether the subject did or did not experience a particular favorable event, such as one-year survival following a surgical procedure. The random assignment of subjects to treatment groups and the (assumed) stochastic independence of outcomes among subjects together guarantee that the total number X^t of subjects in the treatment group who do experience the favorable outcome (which we call a "success") will have the binomial probability distribution with known sample size n^t but unknown success probability p^t; similarly the number X^c of successes in the control group will have the binomial distribution with possibly different parameters n^c and p^c.

2.1. Measures of Treatment Effect

Since the measured outcome was described as "favorable", the treatment will be regarded as effective if p^t exceeds p^c and its effect will be quantified as

$$\varepsilon = g(p^t, p^c) \qquad (2.1)$$

for some function $g(p^t, p^c)$ which vanishes when $p^t{=}p^c$ and is positive when $p^t{>}p^c$. The investigator can choose among many such functions $g(\cdot,\cdot)$, and so among many measures ε of treatment effect.

One such measure of treatment effect is the change in probability of success

$$\mathrm{CP} := p^t - p^c. \qquad (2.2a)$$

With this measure (and the law of large numbers) it is especially easy to predict the increased number of successes if the treatment is given to some number N of subjects-- it is just $N{\times}CP$. An individual patient or physician considering the treatment might be more interested in the relative-risk of failure

$$\mathrm{RR} := (1-p^t)/(1-p^c), \qquad (2.2b)$$

which represents the fractional decrease in the probability $(1-p)$ of the unfavorable outcome. It would be especially meaningful for cases in which the pre-treatment success probability p^c is close to one. In case p^c is close to 0, the fractional-increase in success probability

$$\mathrm{FI} := p^t/p^c \qquad (2.2c)$$

would be more meaningful. Another choice, which (for small effects) is nearly equal to RR when p^c is near 1, to $1/FI$ when p^c is near 0, and to the exponential $e^{4\times CP}$ for moderate p^c, is the odds ratio

$$\mathrm{OR} := \frac{p^t/(1-p^t)}{p^c/(1-p^c)}. \qquad (2.2d)$$

Estimates of the odds ratio are often reported in published accounts of clinical trials or retrospective studies of treatment effect.

It is more convenient to work with a measure of effect which is positive when the treatment is helpful (*i.e.* when $p^t > p^c$) and vanishes when $p^t = p^c$, so we make simple transformations using logarithms where necessary to find:

$$\begin{array}{lrcl} & & & (2.2) \\ \text{Change in Probability:} & \varepsilon_{CP} & := & p^t - p^c \\ \text{Log Relative-Risk:} & \varepsilon_{RR} & := & -\log\left[\dfrac{(1-p^t)}{(1-p^c)}\right] \\ \text{Log Fractional-Increase:} & \varepsilon_{FI} & := & \log(p^t/p^c) \\ \text{Log Odds-Ratio:} & \varepsilon_{OR} & := & \log\left[\dfrac{p^t/(1-p^t)}{p^c/(1-p^c)}\right]. \end{array}$$

The untransformed measures can be recovered as $\mathrm{CP} = \varepsilon_{CP}$, $\mathrm{RR} = e^{-\varepsilon_{RR}}$, $\mathrm{FI} = e^{\varepsilon_{FI}}$, and $\mathrm{OR} = e^{\varepsilon_{OR}}$.

Any of these four choices might be an appropriate way to measure or to report treatment effect; in each case $\varepsilon = 0$ if $p^t = p^c$ (whatever the values of n^t and n^c) and $\varepsilon > 0$ if and only if the treatment improves success probability, *i.e.* $p^t > p^c$. From each measure ε and the value of p^c, any of the other measures can be computed. How is the experimenter to choose among them?

An answer emerges when we consider the problem of combining the evidence about ε from several independent trials, and consider carefully the assumptions we will want to make. For simplicity we will assume that the conditions of the trials were substantially identical except possibly for some inevitable variation in the patient populations (and thus in the success probabilities p_i^t and p_i^c across trials), so that the true treatment effect

$$\varepsilon = g(p_i^t, p_i^c) \qquad (2.3)$$

would not be expected to differ from one study to another. David Eddy (1986) has shown how evidence from trials which *do* differ in substantial ways can sometimes be adjusted for the differences (in order to minimize the effect of between-study variation) and then combined to produce a pooled estimate of treatment effect.

We can easily calculate the joint likelihood function

$$L_{p^t,p^c}(p_i^t, p_i^c \mid X_i^t, X_i^c) = \qquad (2.4)$$

$$\binom{n^t}{X_i^t}(p_i^t)^{X_i^t}(1-p_i^t)^{n_i^t-X_i^t}\binom{n_i^c}{X_i^c}(p_i^c)^{X_i^c}(1-p_i^c)^{n_i^c-X_i^c},$$

which reflects all the experimental evidence from the i^{th} trial about p_i^t and p_i^c. Unfortunately that evidence is not in a form which makes it simple to make inferences about the treatment effect ε.

For any of the four effect measures introduced above in (2.2), the relation (2.3) can be inverted by some function $G(\cdot,\cdot)$ satisfying:

$$p_i^t = G(\varepsilon,\ p_i^c). \tag{2.5}$$

The binomial probability distribution function for X_i^t and X_i^c can now be rewritten as a function of ε and p_i^c, yielding a joint likelihood function for these two parameters:

$$L_{\varepsilon,p^c}(\varepsilon,p_i^c \mid X_i^t,X_i^c) = L_{p^t,p^c}(G(\varepsilon,p_i^c),p_i^c \mid X^t,X^c) \tag{2.6}$$

$$\propto (G(\varepsilon,\ p_i^c))^{X_i^t}\ (1-G(\varepsilon,\ p_i^c))^{n_i^t-X_i^t}\ (p_i^c)^{X_i^c}\ (1-p_i^c)^{n_i^c-X_i^c}$$

Note that the Jacobian

$$J(\varepsilon,\ p^c) = \frac{\partial}{\partial\varepsilon}G(\varepsilon,\ p^c)$$

of the transformation $(p^t,p^c)\rightarrow(\varepsilon,p^c)$ does not enter the formula for the transformed likelihood function, though it does enter the formula for the transformed prior density (and therefor the formula for the transformed posterior density). If we have a prior density function expressing joint uncertainty about p^t and p^c, and wish to transform to one expressing joint uncertainty about ε and p^c, we would calculate it as the product

$$\pi_{\varepsilon,p^c}(\varepsilon,p^c) = \pi_{p^t,p^c}(G(\varepsilon,p^c),p^c)\ J(\varepsilon,\ p^c). \tag{2.7}$$

Of course, it is sometimes more natural to specify the joint prior density function $\pi_{\varepsilon,p^c}(\varepsilon,p^c)$ directly. If it is proper and nondegenerate, it can be written as a product of a marginal and a conditional density function in either of two ways:

$$\pi_{\varepsilon,p^c}(\varepsilon,p^c) = \pi_\varepsilon(\varepsilon)\ \pi_{p^c|\varepsilon}(p^c|\varepsilon) \tag{2.8a}$$

$$= \pi_{p^c}(p^c)\ \pi_{\varepsilon|p^c}(\varepsilon|p^c) \tag{2.8b}$$

If evidence about ε is available from two or more studies, all with the same p^c, we should multiply the product of all the individual likelihood functions times the joint prior density function in order to find a consensus joint posterior, then integrate with respect to p^c and multiply by a normalizing constant c to find a marginal posterior density for ε:

$$\pi_\varepsilon(\varepsilon|X_1,\cdots,X_k) = \tag{2.9}$$
$$c\int_0^1\left\{\prod_{i=1}^{k}L_{\varepsilon,p^c}(G(\varepsilon,p^c),p^c \mid X_i^t,X_i^c)\right\}\pi_{\varepsilon,p^c}(\varepsilon,p^c)\ dp^c.$$

Indeed this possibility of combining evidence from multiple studies is one of the principal reasons Bayesians have for computing likelihood functions (rather than posterior densities) from experiments. It is the product of likelihood functions, and not posterior densities, which must appear between the braces $\{\cdots\}$ in (2.9), to avoid including the prior density function $(k+1)$ times instead of once. For this same reason it is important to include the Jacobian term only when transforming density functions and not when transforming likelihood functions.

Unfortunately, it is frequently the case that the separate trials do *not* share the same p^c, so that (2.9) cannot be used to find a posterior density function for ε. Rather, p_i^c is a nuisance parameter which varies from study to study. In that case it is appropriate to find the posterior density function for ε in each study individually by integrating the product of the individual likelihood function (2.6) and a joint prior density:

$$\pi_\varepsilon(\varepsilon \mid X_i^t,X_i^c) = \tag{2.10a}$$
$$c\int_0^1 L_{\varepsilon,p^c}(G(\varepsilon,p_i^c),p_i^c \mid X_i^t,X_i^c)\pi_{\varepsilon,p^c}(\varepsilon,\ p_i^c)\ dp_i^c.$$

This will give a marginal posterior density function for ε on the basis of the observations $(X_i^t,\ X_i^c)$ from the i^{th} study, but how can we combine them? What we really need are *likelihood* functions for ε, not posterior density functions.

It is not obvious what a "marginal likelihood function for ε" is or how one ought to be computed in the presence of nuisance parameters like p_i^c, but we do know what we would like to be able to do with one—multiply it times a (marginal) prior density function $\pi_\varepsilon(\varepsilon)$ to produce the posterior density function $\pi_{\varepsilon,p^c}(\varepsilon \mid X_i^t,X_i^c)$. Substituting (2.8a) into (2.10a) gives

$$\pi_\varepsilon(\varepsilon \mid X_i^t,X_i^c) = \tag{2.10b}$$
$$c\int_0^1 L_{\varepsilon,p^c}(G(\varepsilon,p_i^c),p_i^c \mid X_i^t,X_i^c)\ \pi_\varepsilon(\varepsilon)\ \pi_{p^c|\varepsilon}(p_i^c|\varepsilon)\ dp_i^c$$

and suggests that we define our marginal likelihood function by the quotient $\pi_\varepsilon(\varepsilon \mid X_i^t,X_i^c)\ /\ \pi_\varepsilon(\varepsilon)$:

$$L_\varepsilon(\varepsilon \mid X_i^t,X_i^c) := \tag{2.11}$$
$$\int_0^1 L_{\varepsilon,p^c}(G(\varepsilon,p_i^c),p_i^c \mid X_i^t,X_i^c)\ \pi_{p^c|\varepsilon}(p_i^c|\varepsilon)\ dp_i^c.$$

Now that we have a likelihood function for each trial, we can combine the evidence from the several trials to find a posterior density function for ε given all the observed data:

$$\pi_\varepsilon(\varepsilon \mid X_1,\cdots,X_k) = \tag{2.12}$$
$$c\ \pi_\varepsilon(\varepsilon)\prod_{i=1}^{k}\left\{\int_0^1 L_{\varepsilon,p^c}(G(\varepsilon,p_i^c),p_i^c \mid X_i^t,X_i^c)\pi_{p^c|\varepsilon}(p_i^c|\varepsilon)dp_i^c\right\}.$$

The use and properties of marginal likelihood functions (2.11) and posterior measures (2.12) are described elsewhere (Wolpert and Berger, 1986).

It is here that an opportunity arises to simplify the analysis. If we can choose an effect measure

$$\varepsilon = g(p_i^t,\ p_i^c)$$

in such a way that ε and p_i^c are *a priori* independent, *i.e.*

$$\pi_{\varepsilon,p^c}(\varepsilon,p^c) = \pi_\varepsilon(\varepsilon)\ \pi_{p^c}(p^c), \tag{2.8c}$$

then the conditional prior $\pi_{p_i^c|\varepsilon}(p_i^c|\varepsilon)$ in (2.11) and (2.12) is just the marginal prior $\pi_{p_i^c}(p_i^c)$, with no functional dependence upon ε. This simplifies the integrals and, moreover, allows us to use a noninformative prior for p_i^c in computing a marginal posterior density function (2.12) for ε. Using a noninformative prior minimizes the influence of any subjective opinion on the analysis.

We will return in Section 3 to the consequences and advantages of using an effect measure independent (under the prior) of p^c, *i.e.* one satisfying (2.8c); we first consider what that independence means in specific examples, and how to achieve it in general.

2.2. The Assumption of Prior Independence

The assumption of prior independence of ε and p_i^c requires the use of an effect measure $\varepsilon = g(p_i^t, p_i^c)$ consistent with that assumption, but simplifies the analysis thereafter by allowing the investigator to ignore prior beliefs and preliminary evidence about the value of p_i^c in each study when searching for evidence from that study about the treatment effect.

Consider once again the four effect measures introduced in (2.2) for binomial trials:

$$
\begin{aligned}
&\text{Change in Probability:} && \varepsilon_{CP} &&= p^t - p^c \\
&\text{Log Relative-Risk:} && \varepsilon_{RR} &&= -\log\left[\frac{(1-p^t)}{(1-p^c)}\right] \\
&\text{Log Fractional-Increase:} && \varepsilon_{FI} &&= \log(p^t/p^c) \\
&\text{Log Odds-Ratio:} && \varepsilon_{OR} &&= \log\left[\frac{p^t/(1-p^t)}{p^c/(1-p^c)}\right].
\end{aligned}
$$

One way to investigate the prior dependence of ε and p^c is to predict how the treatment would act on subpopulations with differing pretreatment success probabilities p^c. Suppose, for example, that the treatment were known to improve the pretreatment success probability of a subpopulation with $p^c = 0.50$ to a post-treatment probability of $p^t = 0.60$. What would be the success probability following treatment for a different subpopulation, one with a pretreatment success probability of only $p^c = 0.25$? Or for a subpopulation with a higher pretreatment success probability of $p^c = 0.75$? The four proposed measures of treatment effect differ in their predictions.

A treatment whose effect is to change the success probability by a fixed amount, and which improves one subpopulation from $p^c = 0.50$ to $p^t = 0.60$, must add 0.10 to p^c for each subpopulation. This would increase p^c=0.25 to p^t=0.35, p^c=0.75 to p^t=0.85, and lead to impossibly high predicted success probabilities for p^c>0.90. Conversely a treatment whose effect is to add a fixed constant to the log-odds, or maintain a fixed odds ratio, and which improves one subpopulation's success odds from 0.50/0.50 = 1.0 to 0.60/0.40 = 1.5, must generate an odds ratio of 1.5 in each subpopulation. This would increase $p^c = 0.25$ (with odds 0.25/0.75 = 1/3) to $p^t = 1/3$ (with odds (1/3)/(2/3) = 0.5 = 1.5×1/3), for a net increase of only 0.0833 in the success probability, and would increase $p^c = 0.75$ (with odds 0.75/0.25 = 3) to $p^t = 9/11$ (with odds (9/11)/(2/11) = 4.5 = 1.5×3) for a net increase of 0.0682. Such an improvement in odds ratio would improve $p^c = 0.01$ or 0.99 only to $p^t = 0.015$ or 0.995, respectively, for an increase of only about 0.005, while $p^c = 0.5$ is increased twenty times as much, a full 0.100. In general a treatment whose effect is measured as an odds ratio would be expected to cause a smaller increase in success probability near the extremes of p^c≈0 or p^c≈1 than near intermediate values such as p^c≈0.5, while one whose effect is measured as a shift in the probability of success would be expected to cause the same size of increase for any p^c. Table 2.1 summarizes some of the predictions of each of these four treatment effect measures, with asterisks (***) indicating an impossibly large prediction.

Effect Measure	Success Probability				
	p^c=.01	p^c=.25	p^c=.50	p^c=.75	p^c=.99
Increased probability	0.110 (+.100)	0.350 (+.100)	0.600 (+.100)	0.850 (+.100)	***
Relative risk	0.208 (+.198)	0.400 (+.150)	0.600 (+.100)	0.800 (+.050)	0.992 (+.002)
Fractional increase	0.012 (+.002)	0.300 (+.050)	0.600 (+.100)	0.900 (+.150)	***
Odds ratio	0.015 (+.005)	0.333 (+.083)	0.600 (+.100)	0.818 (+.068)	0.993 (+.003)

Table 2.1. Measures of Treatment Effect for Binomial Experiments

If we expect that the increase in success probability due to treatment will be smaller for subpopulations with very high or very low initial success probability p^c, then necessarily our prior beliefs about p^c and ε *cannot* be independent if we measure treatment effect as increased success probability $\varepsilon_{lP} := p^t - p^c$. With such a measure the conditional density $\pi_{\varepsilon|p^c}(\varepsilon|p^c)$ would have to be more concentrated near $\varepsilon \approx 0$ for p^c close to 0 or 1 than for p^c close to 0.5, and in particular it must display functional dependence upon p^c. Similarly in this situation p^c and ε could not be independent under the prior distribution if the effect of such a treatment were measured as a decrease in the relative risk or a fractional increase in the success probability; of the four measures introduced in (2.2), only the odds-ratio measure is consistent with a smaller shift $p^t - p^c$ at both extremes than for moderate p^c if p^c and ε are to be independent. Only the relative-risk measure is consistent with a larger shift for small p^c and a smaller shift for large p^c, and only the fractional-increase measure is consistent with the opposite pattern of a smaller shift for small p^c and a larger one for large p^c, if p^c and ε are to be independent.

The example above illustrates how one can approach the problem of choosing a measure of treatment effect in general to assure that, under the prior specification, ε and p^c are stochastically independent. First imagine a subpopulation with moderate population characteristics p_0^c. One at a time consider several possible treatment outcomes p_0^t for the studied treatment on that moderate subpopulation, from the most pessimistic ($p_0^t \ll p_0^c$) to no-effect ($p_0^t \approx p_0^c$) to the most optimistic ($p_0^t \gg p_0^c$). For each, predict the treatment outcome p_i^t for other subpopulations with population characteristics p_i^c varying over the gamut from 0 to 1 (or at least over a subinterval of high prior probability), and try to identify a useful invariant $\varepsilon = g(p^t,p^c)$ (such as $p^t - p^c$, p^t/p^c, $(1-p^t)/(1-p^c)$, *etc.*).

The procedure just described leads to an effect-measure ε whose conditional distribution, given p^c, does not depend on p^c— *i.e.* to an effect measure stochastically independent of p^c.

For convenience in later computations, reparameterize if necessary so that ε=0 denotes no-effect and ε>0 describes a treatment which improves the success probability; in our case, that required taking logarithms or negative logarithms. In the more general setting (described in Wolpert and Berger (1986)) it may also be necessary to introduce a nuisance variable η_i^t describing population attributes in the i^{th} treated population unrelated to treatment effect, so that the treated-group parameter (here $p^t \in (0,1)$, but more generally some parameter θ^t taking values in a parameter space Θ^t) can be written as a function of ε, η_i^t, and the control-group parameter θ^c:

$$\theta^t = G(\varepsilon,\eta_i^t,\theta^c).$$

2.3. Benefits of an Independent Effect Measure

Business leaders and policy makers welcome the opportunity Bayesian analysis affords of explicitly incorporating prior knowledge and subjective beliefs into their analyses, but scientists must scrupulously avoid the appearance of subjectivity. While the (subjective) choice of designs, models, and methods always influences study results regardless of whether the statistical methods used are Bayesian, frequentist, or those of some other school, Bayesian methods have sometimes been criticized on the grounds that their explicit use of prior information precludes objectivity. More recently objective Bayesian methods have been developed which use prior density functions selected on some basis other than subjective beliefs. A number of authors (*e.g.* Berger (1986), Bernardo (1979), Jeffreys (1961), and Box and Tiao (1973)) have recommended methods for selecting prior density functions to meet various objective criteria. These criteria include the preservation of invariance under some change of measurement origin or scale, the maximization of the likelihood function's contribution (and minimization of the prior's contribution) to the Kullback-Leibler measure of information contained in the posterior distribution, and the stationarity of the prior measure under certain infinitesimal deformations. The prior density functions these authors recommend (which do not always coincide) are variously called "noninformative" or "reference" priors.

In general equation (2.12) for the posterior density function for ε calls for integration of each likelihood function with respect to a conditional prior measure on the possible values of p^c. If ε and p^c are stochastically dependent, *i.e.* if the conditional density function for p_i^c given ε does in fact depend on ε, then a noninformative prior density cannot be used in the integral. Independence of ε and p_i^c opens up the possibility of using a noninformative prior density function $\pi_*(dp_i^c)$ in those integrals:

$$\pi_\varepsilon(\varepsilon \mid X_1, \cdots ,X_k) = \qquad (2.12b)$$
$$c\ \pi_\varepsilon(\varepsilon)\prod_{i=1}^{k}\left\{\int_0^1 L_{\varepsilon,p^c}(G(\varepsilon,p_i^c),p_i^c \mid X_i^t,X_i^c)\pi_*(p_i^c)dp_i^c\right\}$$

and hence of avoiding overt subjectivity. While it is true that subjective prior opinion has still played a role (in directing the choice of an effect measure $\varepsilon = g(\cdot)$), that role now enters as part of the unavoidable one of model selection.

Often there are observable patient attributes which might help a clinician make a more specific prediction for a particular patient or subpopulation of patients than is possible for all patients together. For example, in the binomial trial it might be possible to observe p^c

(or evidence bearing on p^c) in individual subjects or subpopulations of subjects. In such situations one would prefer to have a conditional posterior density function $\pi_{p^t|p^c,X}(p^t \mid p^c, X_1, \cdots, X_k)$ expressing the uncertainty remaining about p^t after considering all the evidence from the trials, given the value of p^c, rather than the posterior density function (2.12b) for some abstract measure of treatment effect. If an independent treatment effect measure has been used, then it is possible to derive from (2.12b) the desired conditional posterior density function. For fixed p^c, (2.5) expresses p^t as an explicit function of ε— thus we can just change variables in (2.12b) to calculate this conditional posterior density as

$$\pi_{p^t|p^c,X}(p^t \mid p^c, X_1, \cdots, X_k) = \tag{2.14}$$

$$\pi_\varepsilon(g(p^t,p^c)|X_1, \cdots, X_k)/J(g(p^t,p^c),p^c).$$

With this conditional posterior density for p^t a further change of variables using (2.3) leads to formulas for the conditional posterior densities of any of the four effect measures ε_{CP}, ε_{RR}, ε_{FI}, or ε_{OR} following observation of the available study data. The prior independence of ε and p^c together with the assumption that p^c can vary from study to study (which implies that the studies we have observed offer no evidence about the value of p^c in a later study) allow us to compute and report posterior density functions for any of the four treatment effect measures (or any other which can be expressed in the form (2.3) and (2.5)):

$$\pi_{\varepsilon_{XX}|X_1, \cdots, X_k}(\varepsilon_{XX} \mid X_1, \cdots, X_k) = \tag{2.15}$$

$$\int_0^1 \pi_{p^t|p^c,X}(G_{XX}(\varepsilon_{XX},p^c) \mid p^c, X_1, \cdots, X_k) J_{XX}(\varepsilon_{XX},p^c) \pi_{p^c}(p^c) dp^c.$$

Here ε_{XX}, $G_{XX}(\cdot,\cdot)$, and $J_{XX}(\cdot,\cdot)$ represent the effect size (2.3), inverse function (2.5), and Jacobian for the treatment effect measure chosen for reporting results. Thus the investigator can choose separately the best treatment effect measure for internal computations (presumably, one satisfying the prior independence criterion recommended above) and the best one for the reporting of results (perhaps one with intuitive appeal). Equations (2.12 - 2.15) allow one to derive posterior distributions, either marginally or conditionally given the values of observable patient attributes, for the post-treatment population parameter p^t or for any of the effect measures. There is no reason to choose an intuitively appealing affect measure for internal computations.

3. Summary

Four different measures of treatment effect were introduced, all for assessing improvement (due to treatment) in success probability on the basis of evidence from simple (binomial) randomized controlled trials. The measures were shown to lead to differing predictions for the effect of treatment within subpopulations with characteristics varying from those of the study population.

Objective Bayesian methods for synthesizing the evidence from multiple trials were introduced, and were shown to take an especially simple form if used with a treatment effect measure chosen to satisfy a technical condition: that the parameter vector describing the pre-treatment condition of the treated population be stochastically independent (under the prior probability assessment) of the chosen measure of treatment effect. It was shown how a measure of treatment effect can be selected precisely in order to validate that assumption, and how the results of the research synthesis can be used to provide a quantification of posterior uncertainty about that or any other measure of treatment effect.

In the course of developing the methods a new definition of partial likelihood was introduced, one in which nuisance parameters are removed by integrating with respect to a conditional prior distribution. This notion is developed in detail elsewhere (Wolpert and Berger, 1986). A new method of presenting study results was also introduced, the conditional posterior distribution for treated group population parameters. This too is developed in detail elsewhere (Wolpert and Eddy, 1986).

In research synthesis it is often necessary to make simplifying assumptions in order to reduce the terribly difficult problem of combining evidence from dissimilar trials to a sequence of more manageable problems. Some of these assumptions are about the trials themselves: about the details of the treatment used, the nature of the subjects, or the methods by which subjects were assigned to treatment groups. Others are "technical" in the sense that they concern aspects of the statistical models we construct rather than aspects of the trials themselves. Sometimes it is possible to adapt our proposed model in order to arrange that the technical assumptions be approximately true. This can avoid compromising the validity of an analysis by basing it on assumptions known to be false.

Bayesian statistics, applied with caution, offers a promising methodology for synthesizing the results from clinical trials.

4. References

Berger, J.O. (1986) *Statistical Decision Theory and Bayesian Analysis.* (2nd edn.). Springer-Verlag, New York.

Bernardo, J.M. (1979) Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc.*

Box, G.E.P. and G.C. Tiao (1973) *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, Massachusetts.

Eddy, D.M. (1986) Confidence profiles: a Bayesian method for assessing health technologies. *(appears elsewhere in this issue).*

Jeffreys, H. (1961) *Theory of Probability* (3rd edn.). Oxford University Press, London. 41, 113-147.

Wolpert, R.L. and J.O. Berger (1986) Conditional priors and partial likelihood functions. *(in preparation).*

Wolpert, R.L. and D.M. Eddy (1986) Extensions of the confidence profile method for technology assessment. *(in preparation).*

Comment on Eddy's Confidence Profile Method

David A. Lane, University of Minnesota

David Eddy and his co-workers are in the process of constructing an impressive and potentially very useful technology, whose purpose it is to guide the deliberations of a panel evaluating health practices. They have chosen to build their technology on a Bayesian foundation. The main purpose of this comment is to offer several arguments in support of that choice, which are presented in Section 1 below. In Section 2, I raise some questions about the way in which Eddy implements the Bayesian program and point out some alternative approaches.

1. Bayesian Analysis and Integrative Inference

I claim that Bayesian analysis is the right way to attack the inferential problem that is the focus of this Conference and of Eddy's paper, the problem of integrative inference. To make clear the content of this claim, I will begin by defining the problem of integrative inference and describing what Bayesian analysis is. Then, I will outline the Bayesian approach to integrative inference and discuss its advantages.

The problem of integrative inference: A decision-maker is trying to determine what course of action to take. Which action is appropriate depends on what the consequences of the various possible actions will be. So the decision-maker asks a panel of experts to predict these consequences. For example, he might ask the experts to tell him what risk of cancer is assumed by some particular population if it is exposed to various levels of a particular chemical. The decision-maker expects the experts' answer to reflect all the evidence available to them, and he needs them to use this evidence to produce their best prediction about what will occur as a function of the action he takes, along with some measure of uncertainty about their prediction.

Typically, there are many different streams of evidence that affect the experts' judgement about the consequences they want to predict. Some of this evidence comes from formal studies that directly relate the actions and outcomes of interest. But there are other sources as well, involving different modes of knowing: theory from basic sciences like chemistry and toxicology; data from laboratory studies using animal models or in vitro cell preparations; even hunches based on professional lore and personal experience. This evidence cannot be ignored; it even affects the way the experts think about the meaning of the data obtained from the best designed formal studies, as they try to generalize or modify them to apply to the special circumstances surrounding the actual predictions on which the relevant decision hinges.

The different streams of evidence may point in different directions. The experts have to evaluate the evidentiary significance of each of the streams, and then they must integrate their evaluations to come up with the predictions on which the choice of action depends. How are the experts to proceed? This is the problem of integrative inference.

Bayesian analysis: The purpose of a Bayesian analysis is to measure the analyst's uncertainty about some propositions that are relevant to the problem at hand, in the light of the available evidence. Two attributes distinguish Bayesian analysis from other statistical methodologies. The first is the broad view Bayesian analysis takes of just what constitutes "evidence". All of the modes of knowing -- theory, "hard" data from formal experimentation, observation and experience -- can yield evidence that is incorporated into a Bayesian analysis. Second, in a Bayesian analysis, all uncertainty, from whatever source, is measured in the same scale, that of subjective probability. This makes it possible to use the laws of probability to combine the uncertainty about particular propositions arising from different sources and ultimately to merge different streams of evidence to obtain an overall judgement about the plausibility of the proposition of primary inferential interest. The use of these laws in this way is supported by a normative theory for reasoning in the face of uncertainty, the theory of coherent inference developed by de Finetti (see de Finetti (1974)).

The Bayesian approach to integrative inference: The Bayesian strategy for integrative inference can be summarized as follows. Expert judgement is used to decompose the problem of primary inferential interest into a series of component problems, each of which is more accessible to the knowledge and experience of the expert analysts. Then, the original problem and each of the component problems is formulated in terms of subjective probability evaluations, and the relations between the problems determine corresponding relations that must obtain between the probability evaluations. Next, the component evaluations are carried out, using techniques of direct expert elicitation or model-based Bayesian updating (when appropriate quantitative data are available). Finally, the solutions to the component evaluations are merged according to the laws of probability to yield an answer to the overall problem. Implementations of this strategy can be found in Eddy (1980), where it is applied to the evaluation of effectiveness of cancer screening, and Lane (1987, 1988), where it is used to develop a procedure for causality assessment for adverse drug reactions.

Advantages of the Bayesian approach: There are three important advantages to the Bayesian approach to integrative analysis, compared to alternatives such as global introspection (where the experts list all the relevant factors and sources of information, and then do an implicit mental integration to reach their overall conclusion), qualitative decision algorithms, or frequentist statistical techniques for combining evidence:

(1) It answers the question that the decision-maker asks. Suppose the decision-maker is concerned about the lifetime attributable incidence of cancer, if the individuals in a specified population are exposed to a chemical at a specified level. The output of a Bayesian integrative analysis will be a probability distribution for that incidence. This distribution describes the expert's uncertainty about what that incidence would be if the population were actually exposed at the indicated level, based on all the evidence available to them. The mean of that distribution is their best prediction for what the attributable incidence will be if the decision-maker adopts a course of action that results in the given level of exposure. As such, it is the appropriate quantity to measure the expected frequency of cancer due to exposure for use in a decision analysis to select the best course of action, according to the normative theory of decision-making under uncertainty developed in Savage (1972). The "spread" of the experts' distribution gives the decision-maker information about how sensitive his choice of action is to the residual uncertainty the experts have about their prediction, in the light of all the available evidence. In contrast, frequentist estimates do not take into account information derived from modes of knowing other than formal studies, except through the imprecise process of model specification, and frequentist measures of uncertainty only take into account "sampling variability" and do not regard uncertainty due to model mis-specification. In addition, it is hard to see how to combine the estimates of a unit-free "effect size" that are obtained in frequentist meta-analysis with measures of the value of the appropriate consequences, as is required in formal analysis to determine the best available course of action.

(2) All the available evidence can be incorporated into a Bayesian analysis, in the most appropriate form. In contrast to frequentist statistical methods, expert opinion can enter explicitly into a Bayesian integrative analysis. Moreover, by skillfully decomposing the problem of primary interest, the questions that elicit the experts' opinions can be formulated in such a way that the experts can understand their meaning unambiguously and actually bring the knowledge and experience that constitute their expertise to bear to answer them. In addition, Bayesian analysis can process quantitative data quantitatively, in contrast to nonstatistical approaches to integrative inference like global introspection or qualitative decision algorithms. That expert opinion and quantitative data are both expressed in terms of subjective probability in a Bayesian analysis means that it is straight-forward to merge evidence from these two different sources.

(3) The rules for combining evidence in Bayesian analysis have a normative justification. The normative force of the combination rules means that it is simply inconsistent to disagree with the global conclusions of a Bayesian analysis without finding an appropriate source for the disagreement at a level localized at those questions of theory and observation where expertise actually resides. In essence, anyone disagreeing with the conclusion derived from a Bayesian analysis must believe either that some relevant piece of evidence was not included (and, if he says what it is, the omission can be easily corrected), or that the experts who carried out the analysis were wrong about some specific conclusion they derived from their shared knowledge base (and, if he gives a convincing reason why, the analysis can be appropriately modified)

2. The Confidence Profile Method

In practice, Bayesian analysis is only as good as the methodology that implements it. If the methodology does not ask the user to produce some relevant piece of evidence, or elicits it in such a way that its actual evidentiary content is obscured, then the analysis will not be based on all the available evidence, theoretical possibilities notwithstanding And if assumptions are built into the methodology that describe how the experts ought to feel about the relation between various propositions, then the analysis will reflect those assumptions rather then the experts' actual beliefs.

To what extent does Eddy's Confidence Profile Method succeed in fulfilling the promise of Bayesian integrative analysis? I will discuss three reservations I have about it, listed in increasing order of seriousness. Two factors must be kept in mind in mitigation of these criticial remarks: first, the method is still in the early stages of its development, so it is likely that later versions will improve its performance; second, no technology will ever achieve the normative status of de Finetti's or Savage's theories.

(1) The role that expertise plays in the method is too restrictive. The method takes the formal study as its primary unit of analysis. Expert opinion

is brought to bear qualitatively to create the chain structure that directs the analysis; from then on, its only role is to adjust results obtained from each of the formal studies that the experts regard as relevant to the analysis. However, direct theoretical arguments from basic science can affect expert opinion about the strength of particular links quite independently of any data obtained from a formal study, and personal experience and professional lore, if carefully expressed and evaluated, can supplement or even substitute for formal studies as evidentiary sources for certain kinds of propositions. Neither the paper nor the demonstration of the method presented at the Conference indicate that these types of evidence would enter into the "confidence profile" (or using a more standard and less confusing terminology, the posterior distribution of the outcome of interest).

(2) The method makes many seemingly unwarranted assumptions of independence. For example, consider equation (1), which asserts that the likelihood function for epsilon, based on the results of n "independent" studies, can be factored as the product of n separate likelihood functions, one for each of the studies. Recall what epsilon is: it is the "true effect of the technology in the circumstances of interest". Now, as Eddy recognizes, the "circumstances of interest" are not typically the circumstances in which the n studies were carried out: for example, patient populations can differ with respect to key demographic variables (like age and sex) and severity of the underlying clinical condition, and the way in which the treatment is administered can change from study to study. As a result, it is essential to adjust the results of each study to take such differences into account, and so Eddy's method requires such an adjustment for external validity.

How do the experts carry out this adjustment? Presumably, they have a mental model that describes how they believe the "true effect" depends on the different variables for which they adjust. The experts' sampling distribution for the results of a study depend on the study's circumstances and the experts' views on how circumstances affect outcome; thus, these distributions do not depend just on epsilon, but on the parameters of the adjustment model as well. Conditioning on the parameters of this model and on epsilon, the experts can regard the results of n "physically independent" studies as independent random variables. However, unless they do not believe that study data can affect their opinions about the relation between the "true effect" of a study and the value of the adjustment variables, the experts' marginal distribution for all the study results, just given epsilon, will not factor as the product of the the marginal distribution for each study result, just given epsilon. This is the factorization that is asserted by equation (1). Since it is nearly impossible to imagine that the experts would not change their opinions about how the "true effect" of a health technology depends on such factors as patient age, sex, baseline clinical condition and mode of delivery of the technology, in the light of data from studies that test the technology on particular patient populations, equation (1) is nearly always false.

The same kind of argument applies to many other formulae in Eddy's paper. What this argument implies is that the experts will make their successive adjustments incoherently, since nothing in the method will guarantee that they update their (implicit) adjustment model parameters in the light of the evidence in the studies they sequentially examine. Since, as Eddy argues, these adjustments constitute an essential step in the process of combining evidence (quantitatively as well as qualitatively), this incoherence is a serious deficiency in the method. Making the adjustment models explicit is the only way to correct this deficiency; carrying it out will be a major undertaking.

Three problems have to be solved. First, the increased complexity of the underlying model imposes knowledge representation difficulties: what needs to be updated in the light of information of which type, and how can this updating be carried out in a computationally efficient way? David Spiegelhalter has developed a very promising approach to this question in his work on structures for Bayesian expert systems; see Speigelhalter (1986) for an introduction to this important line of research. Second, the more parameters in a Bayesian model, the higher the dimension of the integration that needs to be carried out to achieve the appropriate marginal distributions. Recent research in approximate methods for high-dimensional integration in the Bayesian context is summarized in Kass, Tierney and Kadane (1988). The third problem is probably the most difficult: how are the experts' ideas about the form of the adjustment model to be elicited and expressed? This is part of the more general problem with Eddy's method discussed below.

3. The method provides no technology to deal with the difficulties involved in eliciting expert opinion and measuring expert uncertainty. The probabilities that appear in de Finetti's normative theory represent an idealized construction: the price that the evaluator would be neutral between buying and selling a ticket worth $1 of the proposition of interest is true and otherwise valueless. This price exists by an easy monotonicity argument, just as the

price that you would just be willing to pay to purchase a new house *exists*; but it can be exceedingly difficult to *determine* this price exactly, especially since the de Finetti transaction is entirely metaphorical, whereas at least you may be required to "put your money where your mouth is" with respect to the house. Thus, the advantages of de Finetti's normative theory for a procedure based on subjective probability depends to a large extent on how the designers of the procedure solve the problem of *measuring* probabilities with precision.

That people encounter serious difficulties when they try to measure their uncertainty about propositions is well-documented in the psychological literature; see for example Kahneman, Slovic and Tversky (1982). Whether or not an expert can provide a meaningful measure of his uncertainty about a proposition depends crucially on the context in which the proposition is to be interpreted and the care with which it is formulated. The most difficult problem that confronts the designers of any procedure based on subjective probability is to frame assessment tasks that are accessible to the knowledge and experience of the experts who must use the procedure. There is no discussion in Eddy's paper about how he has dealt with this problem with respect to the confidence profile method.

Based on my own work with experts on the causality assessment of adverse drug reactions, I am quite skeptical about the ability of experts to assess directly a distribution for such complicated, "global" adjustment variables as Eddy's tau and beta in a way that accurately and coherently incorporates all their relevant beliefs and opinions. Exactly what questions are the experts asked when they assess these distributions? What internal consistency checks are made to show that they understand what these questions mean and that their answers to them are mutually consistent? What happens when experts disagree? Does the method provide any way to probe for the sources of their disagreement and to construct models for the adjustment factors that are based on a pooled knowledge base and command general agreement among those with the relevant expertise? Eddy's method seems to ignore these questions; its technique for handling subjective probability evaluations appears to be just global introspection, with all its faults and pitfalls (see Lane (1984), made quantitative. For a different approach that attempts to determine which assessment tasks are accessible to the relevant experts and to use the rules of probability to decompose unaccessible problems of interest into accessible components, see Lane (1987, 1988).

LITERATURE CITED

De Finetti B (1974). Theory of Probability. John Wiley, New York.

Eddy D (1980). Screening for Cancer: Theory, Analysis, and Design. Prentice Hall, Englewood Cliffs, N.J.

Kahneman D, Slovic P, Tversky A (1982). Judgement under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge.

Kass R, Tierney L, Kadane J (1987). Asymptotics in Bayesian computation. To appear, Bayesian Statistics 3, ed. J Bernardo, M De Groot, D Lindley, A Smith.

Lane D (1984). A probabilist's view of causality assessment. Drug Information Journal, 18, 323-330.

Lane D (1987). Causality assessment for adverse drug reactions: an application of subjective probability to medical decision making. To appear, Statistical Decision Theory and Related Topics, ed. J Berger, S Gupta

Lane D (1988). Subjective probability and causality assessment (with discussion). To appear, Journal of Applied Stochastic Models and Data Analysis.

Savage L (1972). The Foundations of Statistics Second revised edition. Dover, New York.

Spiegelhalter D (1986). Probabilistic reasoning in predictive expert systems. In Uncertainty in Artificial Intelligence, ed. L Kanal, J Lemmer. North-Holland, Amsterdam.

STATISTICAL ISSUES IN THE META-ANALYSIS OF ENVIRONMENTAL STUDIES

Larry Hedges, The University of Chicago

The rapid growth of research literatures in many areas of scientific interest has led to an almost universal desire to find better ways to understand the accumulated research evidence. The use of statistical methods for combining studies or "meta-analysis" has been one response to the problem of extracting summary evidence from a body of related research results. Statistical methods for combining the results of different studies have recently come into wide use in psychology (Smith & Glass, 1977; Glass & Smith, 1979), sociology (see e.g., Crane & Mehard, 1983), and the biomedical sciences (see e.g., Stampfer, 1982). There is a longer tradition of such work in physical sciences such as chemistry (see e.g., Clarke, 1920) and physics (see e.g., Birge, 1932 or Rosenfeld, 1975). Of course, there is a long tradition of research in statistics and agricultural science on combining the results of research studies (see e.g., Tippett, 1931; Fisher, 1932; Pearson, 1932; Cochran, 1937; Yates and Cochran, 1938).

Many different terms have been used to describe the process of combining results from a series of experiments. The terms meta-analysis and quantitative research synthesis are used in social science; the terms overview and pooling of results are used in the biomedical sciences, and the terms review and critical review are often used in the physical sciences. I prefer the terms research synthesis or research review. In each case the general organization of the problem is the same. In the simplest case, each study provides an estimate of a parameter that is believed to be the same across studies, and these estimates from different studies are combined to yield an overall estimate of the parameter. In more complex (and more realistic) cases the estimates from individual studies are used to study the variation of the parameter across studies. For example, if the parameter of interest is a treatment effect then meta-analyses involve using estimates of treatment effects from individual studies to estimate an overall treatment effect or to study its variation across experiments.

It is important to recognize that combining experiments does not mean pooling raw data. Direct pooling of raw data from different experiments may produce misleading results. One example of the misleading consequences of pooling raw data is Simpson's (1954) paradox, which arises when two 2 X 2 tables both show an effect in a given direction, yet a summary table based on pooling the cell counts of the individual tables shows the opposite effect. A real life example is that recent demographic statistics showed that the death rate has gone down in every Illinois age cohort, but the overall death rate is higher. The substantive explanation, of course, is that the population is now older. The point of this example is that the overall pooled statistics do not reflect the results in the individual age cohorts. Similarly, the results of analyses using raw data pooled across studies may actually conceal some aspects of the results of individual studies.

It is perhaps interesting to note that the term meta-analysis arose as an attempt to describe an activity that was an analysis of the results of the analyses from individual studies (Glass, 1976). The assumption is that the relevant result of the statistical analysis of an individual study is a parameter estimate. The information from different studies is combined via their parameter estimates. If these estimates are sufficient statistics then of course no information is lost in combining information via these statistics.

1.0 Why Is It Desirable to Combine the Results of Studies?

One of the first questions that arises in connection with combining research results is why is it desirable to do so. The answer depends to some extent on the use that is to be made of the research review. Policy decisions and many scientific decisions frequently require information about a phenomenon under a wide range of conditions. For example, these decisions may involve questions about the probable effect of a treatment under a specific, even eccentric set of conditions. Research reviews are perhaps most helpful to inform decisions that involve general conclusions about a typical range of situations. They are likely to be less helpful in identifying precisely what to expect in a very specific or a very eccentric situation. There are, however, several general reasons to expect that combining evidence will be useful.

1.1 Synthesis Provides Robust Evidence.

Even very similar studies differ in their experimental conditions and in the details of their execution. Obviously, planned and plausibly relevant differences in study design or procedure may result in differences in study results.

There are also many subtle, unplanned, and often unrecognized differences between studies that often lead to variation in study results. Seemingly irrelevant differences in experimental conditions, procedures, or measurement methods quite frequently lead to substantial variation in study results. Even large single studies involve a limited set of experimental conditions and context which reduces the generalizability of their findings.

Syntheses based on several studies draw evidence from across contexts to provide conclusions that are more robust to variations in experimental context and thus more useful for broad policy decisions.

1.2 Syntheses May Utilize a Less Biased "Sample" of the Evidence.

The selection of evidence on which to base a policy decision is essentially a sampling process. Haphazard or uncontrolled sampling of evidence can lead to very substantial biases. Unfortunately there are often sharply conflicting interests in the policy making context which may have vested interest in drawing attention to the particular subsets of the research evidence that support their viewpoint. Research results are often part of the rhetoric of competing interests. Thus there is a natural tendency for competing interests to emphasize different parts of the total body of research evidence (different studies) which most closely corresponds to their own beliefs. That is, different interests have a tendency to emphasize biased subsamples of the total body of research evidence. One crucial aspect of quantitative systhesis of research is arriving at a minimally biased sample of research studies in which to base inferences. Carefully operationalized procedures for selecting research evidence can reduce both bias and the appearance of bias in study selection.

1.3 Synthesis Formalizes What Policy Makers Must Do Anyway.

Policy makers are frequently required to make decisions even though they are faced with many studies yielding possibly inconclusive or conflicting findings. That is, policy makers will derive general conclusions. The only question is whether they do so via formal, quantitative means or informal, intuitive means. One of the difficulties in relying on the intuitive procedures for combining research results is that the intuition of many people (even sophisticated people) is horrendously bad (see Hedges & Olkin, 1985). Procedures that seem intuitively sensible are often highly misleading. For example, there is a tendency to look for the prepondenance of evidence by asking what proportion of the studies actually found a statistically significant effect. Examination of such a procedure will demonstrate that when effects of interest are small, this procedure not only has a very poor chance of detecting a real treatment effect, but its properties do not improve as the amount of evidence (the number of studies) increases (see Hedges and Olkin, 1985, pages 48-52). The fundamental problem is that research results, whether they are expressed as estimates, the outcomes of significance tests, or p-values, have a substantial scholastic component. It is simply difficult to find the structure in a series of scholastic observations without some sort of formal methods. Informal methods either ignore a great deal of information or use it suboptimally. Moreover, it is difficult to deduce the properties of informal methods of combining research results. Thus even if they were adequate, it would be difficult to produce a convincing demonstration that this was the case.

1.4 Synthesis Increases Statistical Power.

One of the most obvious advantages of quantitative research synthesis is that the pooling of information from different studies increases the statistical power of hypothesis tests (e.g., for the treatment effect) and decreases standard error of estimates of the experimental effect. Thus the evidence from several studies that are marginal or submarginal in terms of statistical power but are otherwise well done might be combined to yield a relatively powerful test for the existence of a treatment effect.

1.5 Synthesis Provides Formal Standards of Rigor for the Process of Accumulating Evidence from Different Research Studies.

The generation of scientific knowledge begins with the conduct of the individual experiments to generate empirical evidence. A single experiment seldom stands alone, however. The generation of new scientific knowledge almost invariably involves the synthesis of evidence from several replicated studies. Evidence from relevant experiments becomes part of the scientific knowledge-base only after it has been suitably synthesized and interpreted. Every scientist learns rules of methodology or procedure that are designed to insure the validity of original research studies. Because the combination of evidence across studies is just as important to the generation of scientific knowledge as the combination of evidence (from different observations) within studies, rigorous procedure in original research. Moreover, procedural rigor serves the same purpose in both contexts: It protects the validity of conclusions from potential sources of bias. In fact, the parallels between procedure in original research and research synthesis are usually used as the basis for examining rigor in research synthesis.

This paper is an examination of statistical issues in research synthesis. I am using a broad definition of statistical issues which assumes statisticians have a contribution to make in every stage of a research enterprise. The convenience, the treatment that follows is organized into a sequence of generic stages that apply equally well to any original research study or to any research synthesis.

2.0 Issues in Problem Formulation.

Problem formulation is often conceptualized as the first step in any original research study or research review (Cooper, 1984; Light & Pillemer, 1984). It involves formulating the precise questions to be answered by the review. One aspect of formulating questions is deciding whether the purpose of the review is confirmatory (hypothesis testing) or exploratory (hypothesis generating).

A second aspect of problem formulation concerns decisions about when studies are similar enough to combine. That is, deciding whether treatments, controls, experimental procedure, and study outcome measures are comparable enough to be considered. Philosophers of science have provided a conceptualization that is helpful in thinking about this problem. They distinguish the theoretical or conceptual variables about which knowledge is sought (called constructs) from the actual examples of these variables that appear in studies (called operations).

For example, we may want to know if a particular method of teaching mathematics leads to better mathematical problem solving. To find out, a comparative study is conducted in which students are randomly assigned to teachers, some of whom use the new method. The students are then given a problem solving test to determine which group of students were better at mathematical problem solving. The exact conceptualization of mathematical problem solving is a construct. The particular test used to measure problem solving is an operation corresponding to that construct.

Similarly, the particular teaching method as defined conceptually is a construct, while the behavior of a particular teacher trying to implement that teaching method is an operation. The point here is that even when studies share the same constructs, they are sure to differ in the operations that correspond to those constructs.

Thus defining questions precisely in a research review involves deciding on the constructs of independent variables, study characteristics and outcomes that are appropriate for the questions addressed by the review and deciding on the operations that will be regarded as corresponding to the constructs. That is, the reviewer must develop both construct definitions and a set of rules for deciding which concrete instances (of treatments, controls, or measures) correspond to those constructs.

Although the questions of interest might seem completely self-evident in a review of related clinical trials, a little reflection may convince you that there are subtleties in formulating precise questions. For example, consider clinical trials in which the outcome observed is the death rate. At first, the situation seems completely clearcut, but there are subtleties. Should deaths from all causes be included in the death rate or only deaths related to the disease under treatment? If the latter approach is used, how are deaths related to side effects of the treatment to be counted? If there is follow-up after different intervals, which intervals should be used? Should unanticipated or data defined variables (such as "sudden death") be used? Careful thinking about the problem under review usually leads to similar issues which require consideration.

2.1 Selecting Constructs and Operations.

One of the potential problems of meta-analysis or quantitative research syntheses is that they may combine incommensurable evidence. Some meta-analysis have been criticized for combining "apples and oranges." This is essentially a criticism of the breadth of constructs and operations chosen. In one sense breadth of constructs and operations chosen must reflect the breadth of the question addressed by the review. The issue is complicated by the fact that constructs and operations chosen must reflect the breadth of the questions addressed by the review. The issue is complicated by the fact that constucts and operations are often distinguished more narrowly by the reviewer than may be reflected in the final presentation of results. Thus the issue of constructs and operations are to be included in the review and then what constructs and operations are to be distinguished in the data analysis of the review, and finally which constructs and operations are presented in the results of the review.

Meta-analyses have tended to use rather broad constructs and operations in their presentation of results (Cook & Leviton, 1980). This may have resulted from the arguments of Glass and his associates (Glass, McGaw, & Smith, 1981) who urged meta-analysts to seek general conclusions. It may also be a consequence of the ease with which quantitative methods can analyze data from large numbers of studies (Cooper & Arkin, 1981). It is important to recognize however that while broad questions necessitate the inclusion of studies with a range of constructs and operations, they need not inhibit the meta-analyst from distinguishing variations of these constructs in the data analysis and in presentation of results.

2.1.1 Broad Versus Narrow Constructs.

The advantage of broad constructs and operations is that they may support broad generalizations. Because they are maximally inclusive, broad constructs and operations obviate most arguments about studies that should have been included, but were not.

However, the uncritical use of very broad constructs in meta-analysis is problematic. Analyses based on operationalization of broad constructs are vulnerable to the criticism that overly broad choices of construct obscure important differences among the narrower constructs subsumed therein. For example Presby (1978) argued that the broad categories of therapies used by Smith and Glass (1977) in their review of studies of the effectiveness of psychotherapy obscured important differences between therapies and their effectiveness. A similar argument may be made about the breadth of outcome constructs. Moreover, empirical data from research synthesis sometimes confirm the truth of these arguments.

Perhaps the most successful applications of broad or multiple constructs in meta-analysis are those that may include broad constructs in the review but distinguish narrower constructs in the data analysis and presentation of results. This permits the reviewer to examine variations in the pattern of results as a function on construct definition. It also permits separate analyses to be carried out for each narrow construct (see e.g., Cooper, 1979, Linn & Peterson, 1985, Eagly & Carli, 1981; Thomas & French, 1985). A combined analysis across constructs may be carried out where appropriate or distinct analyses for the separate constructs may be presented.

2.1.2 Broad Versus Narrow Operations for Constructs.

Another issue of breadth arises at the level of operationalization of constructs. The reviewer will always have to admit several different operations for any given construct. Treatments will not be implemented identically in all studies and different studies will measure the outcome construct in different ways. Thus, the reviewers must judge whether each operation is a legitimate representation of the corresponding construct. This involves obtaining as much information as possible about the treatment actually implemented and the outcome actually used in each study. This may involve the use of secondary sources such as technical reports, general descriptions of treatment implementations, test reviews, or published tests.

In spite of the difficulty they may present to the reviewer, multiple operations can enhance the confidence in relationships between constructs if the analogous relationships between operations hold under a variety of different (and each imperfect) operations (Campbell, 1969). However, increased confidence comes from multiple operations only when the different operations are in fact more related to the desired construct than to some other construct (see Webb, Campbell, Sechrest, & Grove, 1981 for a discussion of multiple operationism). Thus although multiple operations can lead to increased confidence through "triangulation" of evidence, the indiscriminate use of broad operations can also contribute to invalidity of results via confoundings of one construct with another (see Cooper, 1984).

2.2 Exploratory versus Confirmation Reviews.

A crucial aspect of problem formulation is distinguishing whether the purpose of the review is to test a small number of reasonably well-defined hypotheses, or to generate new hypotheses. Obviously new hypotheses (even new variables) arise in the course of meta-analyses, just as in any scientific activity. The critical issue is to distinguish the clearly a priori hypotheses from those that are suggested by the data. This distinction has implications for the choice of statistical analysis procedures used in the meta-analysis and for interpretation of results. Most statistical tests calculate levels of statistical significance assuming that the hypothesis is a priori and is tested in isolation. When statistical tests are suggested by the data the usual procedures for assessing statistical significance are likely to be misleading. Similarly, when many statistical analyses are conducted on the same data, the usual significance levels will not reflect the chance of making at least one Type I error in the collection of tests (the simultaneous significance level). Thus when conducting many tests in an exploratory mode there is a tendency to "capitalize on chance."

One method of dealing with the problem of exploratory analysis in research reviews is to use statistical methods that are specifically designed for exploratory analysis such as clustering methods (Hedges & Olkin, 1983, 1985). Another alternative is to adjust the significance level to reflect the fact that many tests are conducted on the same data (Hedges & Olkin, 1985). The problem with this and all other simultaneous procedures is that they reduce the power of statistical tests and the effect is dramatic when many tests are conducted simultaneously.

Another alternative is the use of procedures that do not involve statistical significance. The simplest procedures are simply descriptive statistics. Graphical procedures such as Light and Pillemer's (1984) funnel diagrams, Hedges and Olkin's (1985) confidence interval plots, or many of the graphical ideas presented by Tukey (1977) may also be helpful.

A third alternative is to randomly divide the data into two subsets. The first subset is used to generate hypotheses whose statistical significance is then evaluated (cross validated) on the second subset (Light & Pillemer, 1984).

3.0 Issues in Data Collection.

Data collection in meta-analysis consists of assembling a collection of research studies and extracting quantitative indices of study characteristics and of effect magnitude (or relationship between variables). The former is largely a problem of selecting studies that may contain information relevant to the specific questions addressed in the review. It is largely a sampling process. The latter is a problem of obtaining quantitative representations of the measures of effect magnitude and the other characteristics of studies that are relevant to the specific questions addressed by the review. This is essentially a measurement process similar to other complex tasks or judgments that researchers are sometimes required to make in other research contexts. The standard psychological measurement procedures for ensuring the reliability and validity of such ratings or judgments are as appropriate in meta-analysis as in original research (Rosenthal, 1984; Stock, Okun, Haring, Miller, Kinney, & Ceurvorst, 1982).

3.1 Sampling in Meta-analysis.

The problem of assembling a collection of studies is often viewed as a sampling problem: The problem of obtaining a representative sample of all studies that have actually been conducted. Because the adequacy of samples necessarily determines the range of valid generalizations that are possible, the procedures used to locate studies in meta-analysis have been regarded as crucially important. Much of the discussion on sampling in meta-analysis (e.g., Cooper, 1984; Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982; Rosenthal, 1984), concentrates on the problem of obtaining a representative or exhaustive sample of the studies that have actually been conducted. However, this is not the only or even the most crucial aspect of sampling in meta-analysis. Another, equally important sampling question is whether the samples of subjects and treatments in the individual studies are representative of the subject populations and treatment populations of interest.

The importance of representative sampling of subjects is obvious. For example, studies of the effects of psychotherapy on college students who do not have psychological problems may not be relevant to the determination of the effects of psychotherapy on patients who have real psychological problems. The importance of representative sampling of treatments is perhaps more subtle. The question is whether the treatments which occur in studies are representative of the situations about which the reviewer seeks knowledge (Bracht & Glass, 1968). A representative sample of studies, each of which involves a nonrepresentative sample of subjects or treatments, brings us no closer to the truth about the subject or treatments that we care about.

Thus there are two levels of sampling to be concerned about in meta-analysis. One level concerns the mechanism for selecting the sample of studies that are used in the review. The other concerns the mechanism used within studies for selecting individual replication. The situation is much like that of two-stage samples in sample surveys. The reviewer samples clusters or secondary sampling units first; then the individual subjects or primary sampling units are sampled from the clusters.

Strategies for obtaining respresentative or exhaustive samples of studies have been discussed by Glass, McGaw, and Smith (1981) and Cooper (1984). The problem of obtaining representative samples of subjects and treatments is constrained by the sampling of studies and consequently is not under the complete control of the reviewer. The reviewer can, however, present descriptions of the samples of subjects and treatments and examine the relationship between characteristics of these samples and study outcomes. Such assessments of the representativeness of treatments and subjects are obviously crucial in evaluation of the studies on which the review is based.

3.2 Missing Data in Meta-analysis.

Missing data is a problem that plagues many forms of applied research. Survey researchers are well aware that the best sampling design is ineffective if the information sought cannot be extracted from the units that are sampled. Of course missing data is not a substantial problem if it is "missing at random," that is, if the missing information is essentially a random sample of all the information available (Rubin, 1976). Unfortunately there is usually very little reason to believe that missing data in meta-analysis is missing at random. On the contrary, it is often easier to argue that the causes of the missing data are systematically related to effect size or to important characteristics of studies. When this is true, missing data poses a serious threat to the validity of conclusions in meta-analysis. The specific cases of missing data on study outcome and missing data on study characteristics are considered separately.

3.2.1 Missing Data on Study Outcome.

Studies (such as single case studies) that do not use statistical analyses are one source of missing data on study outcome. Other studies use statistics but do not provide enough statistical information to allow the calculation of an estimate of the appropriate outcome parameter. Sometimes this is a consequence of failure to report relevant statistics. More often it is a consequence of the researcher's use of a complex design that makes difficult or impossible the construction of a parameter estimate that is completely comparable to those of other studies. Unfortunately both the sparse reporting of statistics and the use of complex designs are plausibly related to study outcomes. Both result at least in part from the editorial policies of some journals which discourage reporting of all but the most essential statistics. Perhaps the most pernicious sources of missing data are studies which selectively report statistical information. Such studies typically report only information on the effects that are statistically significant, exhibiting what has been called reporting bias (Hedges, 1984). Missing effects can lead to very serious biases, identical to those caused by selective publication which are discussed in the section on publication bias.

One strategy for dealing with incomplete effect size data is to ignore the problem. This is almost certainly a bad strategy. If nothing else, such a strategy reduces the credibility of the meta-analysis because the presence of at least some missing data is obvious to knowledgeable readers. Another problematic strategy for handling missing effect size data is to replace all of the missing values by the same imputed value (usually zero). Although this strategy usually leads to a conservative (often extremely conservative) estimate of the overall average effect size, it creates serious problems in study characteristics to effect size. A better strategy is to extract from the study any available information about the outcome of the study. The direction (sign) of the effect can often be deduced even when an effect size cannot be calculated. A tabulation of these directions of effects can therefore be used to supplement the effect size analysis (e.g., Giaconia & Hedges 1982; Crain & Mahard, 1983). Such a tabulation can even be used to derive a parametric estimate of effect (Hedges & Olkin, 1980, 1985).

Perhaps the best strategy to deal with missing data on study outcomes is the use of the many analytic strategies that have been developed for handling missing data in sample surveys (Madow, Nisselson, & Olkin 1983; Madow, Olkin, & Rubin, 1983; Madow & Olkin, 1983). Generally these strategies involve using the available information (including study characteristics) to estimate the structure of the study outcome data and the relationships among study characteristics and study outcome. They can also be used to study the sensitivity of conclusions to the possible effects of missing data. Although these strategies have much to recommend them they have only rarely been used in meta-analysis.

3.2.2 Missing Data on Study Characteristics.

Another less obvious form of missing data is missing data on study characteristics which results from incompletely detailed descriptions of the treatment, controls, experimental procedure, or the outcome measures. In fact, the generally sketchy descriptions of studies in the published literature often constrain the degree of specificity possible in schemes used to code between study differences.

The problem of missing data about study characteristics is related to the problem of breadth of constructs and operations for study characteristics. Coding schemes that use a high degree of detail (and have higher fidelity) generally result in a greater degree of missing data. Consequently, vague study characteristics are often coded on all studies or more specific characteristics are coded on a relatively few studies (see Orwin & Cordray, 1985). Neither procedure alone seems to inspire confidence among some readers of the meta-analysis.

One strategy for dealing with missing information about study characteristics is to have two levels of specificity: a broad level which can be coded for nearly all studies and a narrower level which can be coded for only a subset of the studies. This strategy may be useful if suitable care is exercised in describing the differences between the entire collection of studies and the smaller number studies permitting the more specific analysis. A more elegant solution is the use of the more refined methods for handling missing data on study characteristics are little used but deserve more attention. One is the collection of relevant information from other sources such as technical reports, other more general descriptive reports on a program, test reviews or articles that describe a program, treatment, or measurement method. The appropriate references are often published in research reports. A second and often neglected source of information is the direct collection of new data. For example in a meta-analysis of sex differences in helping behaviors, Eagly and Crowley (1986) surveyed a new sample of subjects to determine the degree of perceived danger in the helping situations examined in the studies. This rating of degree of perceived danger to the helper was a valuable factor in explaining the variability of results across studies.

3.3 Publication Bias.

An important axiom of survey sample design is that an excellent sample design cannot guarantee a representative sample if it is drawn from an incomplete enumeration of the population. The analogue in meta-analysis is that an apparently good sampling plan may be thwarted by applying the plan to an incomplete and unrepresentative subset of the studies that were actually conducted.

The published literature is particularly susceptible to the claim that it is unrepresentative of all studies that may have been conducted (the so-called publication bias problem). There is considerable empirical evidence that the published literature contains fewer statistically insignificant results than would be expected from the complete collection of all studies actually conducted (Bozarth & Roberts, 1972; Hedges, 1984b; Sterling, 1959). There is also direct evidence that journal editors and reviewers intentionally include statistical significance among their criteria for selecting manuscripts for publication (Bakan, 1966; Greenwald, 1975; Melton, 1962). The tendency of the published literature to over-represent statistically significant findings leads to biased overestimates of effect magnitudes from published literature (Lane & Dunlap, 1978; Hedges, 1984b), a phenomenon that was confirmed empirically by Smith's (1980a) study of ten meta-analyses, each of which presented average effect size estimates for both published and unpublished sources.

Reporting bias is related to publication bias based on statistical significance. Reporting bias creates missing data when researchers fail to report the details of results of some statistical analyses, such as those that do not yield statistically significant results. The effect of reporting bias is identical to that of publication bias: some effect magnitude estimates are unavailable (e.g., those that correspond to statistically insignificant results).

Publication or reporting bias may not always be severe enough to invalidate meta-analyses based solely on published articles (see Light and Pillemer, 1984; Hedges, 1984b). Theoretical analysis of the potential effects of publication bias showed that even when nonsignificant results are never published (the most severe form of publication bias), the effect on estimation of effect size may not be large unless both the within study sample sizes and the underlying effect size are small. However, if either the sample sizes in the studies or the underlying effect sizes are small, the effect on estimation can be substantial.

The possibility that publication or reporting bias may inflate effect size estimates suggests that reviewers may want to consider investigating its possible impact. One method is to compare the effect size estimates derived from published (e.g., books, journal articles) and unpublished sources (e.g., conference presentations, contract reports, or doctoral dissertations). Such comparisons however are often problematic because the source of the study is often confounded with many other study characteristics. An alternative procedure is to use statistical corrections for estimation of effect size under publication bias. This corresponds to modeling the sampling of studies as involving a censoring or truncation mechanism. If these corrections produce a negligible effect, this suggests that publication and reporting bias are negligible.

There have been relatively few detailed statistical analyses of the existence and magnitude of publication and reporting bias. Such studies are badly needed as are refinements of statistical analysis tools to handle less extreme and more realistic censoring models than those considered thus far.

4.0 Issues in Data Evaluation

Data evaluation in meta-analysis is the process of critical examination of the corpus of information collected, to determine which study results are expected to yield reliable information. Judgments of study quality are the principal method of data evaluation. A second aspect of data evaluation is the use of empirical methods to detect outliers or influential data points. When properly applied, empirical methods have uses in both meta-analysis (Hedges & Olkin, 1985) and in primary research (Barnett and Lewis, 1978; Hawkins, 1980).

Meta-analysts and other reviewers of research have sometimes used a single binary (high/low) judgment which may be useful for some purposes such as deciding which studies to exclude from the review. It is seldom advisable to make such judgments directly. The reason is that different researchers do not always agree on which studies are of high quality. Empirical research suggests that direct ratings of study quality have very low reliability (see Orwin & Cordray, 1985). Consequently, most meta-analysts at least initially characterize study quality by using multiple criteria. One approach to criteria for study quality is the threats-to-validity approach, in which each study is rated according to the presence or absence of some general threats to validity such as those presented by Campbell and Stanley (1963) or Cook and Campbell (1979). A second approach is the methods-description approach (Cooper, 1984) in which the reviewer exhaustively codes the stated characteristics of each study's method. A third approach to assessing study quality is a combination of the first two approaches involving coding of the characteristics of study methodology and assessing threats to validity that may not be reflected in the characteristics of study methods (Cooper, 1984).

Another source of information in data evaluation comes from data analyses themselves. It often happens that one or more observations (estimates of effect magnitude) fail to fit the pattern of the other observations. That is, one or more of the data points fail to conform to the same model as do the other observations. These deviant observations or outliers may be the result of studies or situations in which the treatment is exceptionally powerful or exceptionally weak. In some cases a careful examination of details of study design or procedures suggests plausible reasons for the exceptional treatment effect.

Although statistical methods may be used to detect outliers in meta-analysis (Hedges & Olkin, 1985), the question of what to do about them cannot always be resolved so easily. Outliers that result from detectable (and remediable) errors in computation should of course be replaced by estimates based on the correct calculations. When they are based on suspicious data then a cautious data analyst might want to delete, or at least consider separately, such suspicious observations.

The most difficult problem arises when examination of the outlying studies reveals no obvious reason why their effects sizes should differ from the rest. The analysis of data containing some observations that are outliers (in the sense of not conforming to the same model at the other studies) is a complicated task. It invariably requires the use of good judgment and the making of decisions that are in some sense, compromises. There are cases (Rocke, Downs, & Rocke, 1982; Stigler, 1977; Tukey, 1977) where setting aside a small proportion of the data (certainly less than 15-20 percent) has some advantages. If nearly all the data can be modeled in a simple, straightforward way it is certainly preferable to do so, even at the risk of requiring elaborate descriptions of the studies that are set aside.

Studies that are set aside should not be ignored; often these studies reveal patterns that are interesting in and of themselves. Occasionally these deviant studies share a common characteristic that suggests an interesting direction for future research. One of the reasons it is preferable to use a model which includes most of the data is that the results of studies that are identified statistically as outliers often do not deviate enough to disagree with the substantive result of the model. That is, an effect size estimate may exhibit a statistically significant difference from those of other studies, yet fail to differ from the rest to an extent that would make a practical or substantive difference. However, it is crucial that all data be reported and that any deleted data be clearly noted.

5.0 Data Analysis and Interpretation.

Data analysis and interpretation are the heart of the meta-analysis and have a long history in statistics and the physical sciences. Two distinctly different directions have been taken for combining evidence from different studies in agriculture almost from the very beginning of statistical analysis in that area. One approach relies on testing for statistical significance of combined results across studies, and the other relies on estimating treatment effects across studies. Both methods date from as early as the 1930's (and perhaps earlier) and continue to generate interest among the statistical research community to the present day.

Testing for the statistical significance of combined data from agricultural experiments is perhaps the older of the two traditions. One of the first proposals for a test of the statistical significance of combined results (now called testing the minimum p or Tippett method) was given by L.H.C. Tippett in 1931. Soon afterwards, R.A. Fisher (1932) proposed a method for combining statistical significance, or p-values, across studies. Karl Pearson (1933) independently derived the same method shortly thereafter, and the method variously called Fisher's method or Pearson's method was established. Research on tests of the significance of combined results has flourished since that time, and now well over 100 papers in the statistical literature have been devoted to such tests. A review of this literature with special reference to meta-analysis is given in Hedges and Olkin (1985).

Tests of the significance of combined results are sometimes called omnibus or nonparametric tests because these tests do not depend on the type of data or the statistical distribution of those data. Instead, tests of the statistical significance of combined results rely only on the fact that p-values are uniformly distributed between zero and unity. Although omnibus tests have a strong appeal in that they can be applied universally, they suffer from an inability to provide estimates of the magnitude of the effects being considered. Thus, omnibus tests do not tell the experimenter how much of an effect a treatment has. Consequently omnibus tests are of limited utility in most research reviews.

In order to determine the magnitude of the effect of an agricultural, a second approach was developed which involved combining numerical estimates of treatment effects. One of the early papers on the subject (Cochran, 1937) appeared a few years after the first papers on omnibus procedures. Additional work in this tradition appeared shortly thereafter (e.g., Yates & Cochran, 1938; Cochran, 1943). It is also interesting to note that work on statistical methods for combining estimates from different experiments in physics dates from the same era (Birge, 1932).

6.0 Combined Significance Tests.

This section outlines some of the methods used for combined significance testing.

Consider a collection of k independent studies characterized by parameters $\theta_1, \ldots, \theta_k$, such as means, mean differences, or correlations. Assume further that the ith study produces a test statistic T_i to be used to test the null hypothesis

$$H_{0i}: \theta_i = 0, \; i=1, \ldots, k,$$

where large values of the test statistic lead to rejection of the null hypothesis. The hypothesis $H_{0i}, \ldots, H_{0k}$ need not have the same substantive meaning, and similarly, the statistics $T_1, \ldots, T_k$ need not be of related form. The omnibus null hypothesis H_0 is that none of the effects is significant, that is, that all the O's are zero:

$$H_0: \theta_1 = \theta_2 = \ldots = \theta_k = 0$$

Note that the composite hypothesis H_0 holds only if each of the subhypotheses $H_{01}, \ldots, H_{0k}$ holds.

The one-tailed p-value for the ith study is

$$P_i = \text{Prob}\,\{T_i \geq t_{i0}\}\,. \qquad (1)$$

where t_{i0} is the value of the statistic actually obtained (the sample realization of T_i) in the ith study. If H_{0i} is true, then P_i is uniformly distributed i the interval [0,1].

6.1 The Minimum p Method.

The first test of the significance of combined results was proposed by Tippett (1931), who pointed out that if $p_1, \ldots, p_k$ are independent p-values (from continuous test statistics), then each has a uniform distribution under H_0. Therefore, if $p_{[1]}$ is the minimum of $p_1, \ldots, p_k$, a test of H_0 at significance level is obtained by comparing $p_{[1]}$ with $1-(1-\alpha)^{1/k}$, so that the test procedure is to

$$\text{reject } H_0 \text{ if } p_{[1]} \leq 1-(1-\alpha)^{1/k} \qquad (2)$$

6.2 The Inverse Chi-Square Method.

Perhaps the most widely used combination procedure is that of Fisher (1932). Given k independent studies and the p-values $p_1, \ldots, p_k$, Fisher's procedure uses the product, $p_1, p_2, \ldots, p_k$, to combine the p-values. He made use of a connection between the uniform distribution and the chi-square distribution, namely, that if U has a uniform distribution then -2 log U has a chi-square distribution with two degrees of freedom. Consequently, when H_{0i} is true, -2 log p_i has a chi-square distribution with two degrees of freedom. Because the sum of independent chi-square variables has a chi-square distribution, we have the very simple and elegant fact that if H_0 is true, then

$$-2 \log(p_1, p_2, \ldots, p_k) = -2 \log p_1 - \ldots -2 \log p_k \qquad (3)$$

has a chi-square distribution with 2k degrees of freedom. Because of this fact, no special tables are needed for the Fisher method. The test procedure becomes

$$\text{reject } H_0 \text{ if } P = -2 \sum_{i=1}^{k} \log p_i \geq C,$$

where the critical value C is obtained from the upper tail of the chi-square distribution with 2k degrees of freedom.

6.3 The Inverse Normal Method.

Another procedure for combining p-values is the inverse normal method proposed by Stouffer, Suchman, Devinney, Star and Williams (1949). This procedure involves transforming each p-value to the corresponding normal score, and then "averaging." More specifically, define z_i by $p_i = \Phi(z_i)$, where $\Phi(x)$ is the standard normal cumulative distribution function. When H_0 is true, the statistic

$$Z = \frac{z_1 + \ldots + z_k}{\sqrt{k}} = \frac{\Phi^{-1}(p_1) + \ldots \Phi^{-1}(p_k)}{\sqrt{k}} \tag{4}$$

has the standard normal distribution. Hence we reject H_0 whenever Z exceeds the appropriate critical value of the standard normal distribution.

6.4 The Logit Method.

Yet another method for combining k independent p-values $p_1, \ldots, p_k$ was suggested by George (1977) and investigated by Mudholkar and George (1979). Transform each p-value into a logit, $\log[p/(1 - p)]$, and then combine the logits via the statistic

$$L = \log \frac{p_1}{1 - p_1} + \ldots + \log \frac{p_1}{1 - p_k}. \tag{5}$$

The exact distribution of L is not simple, but when H_0 is true, George and Mudholkar (1977) show that the distribution of L (except for a constant) can be closely approximated by Student's t-distribution with 5k + 4 degrees of freedom. Therefore, the test procedure using the logit statistic is

$$\text{reject } H_0 \text{ if } L^* = |L| \sqrt{(0.3)(5k+4)/k(5k+2)} > C \tag{6}$$

where the critical value C is obtained from the t-distribution with $5k_2$+ 4 degrees of freedom. [The term 0.3 in (6) is more accurately given by $3/\pi^2$.

For large values of k, $\sqrt{3(5k+4)/\pi^2(5k+2)} \doteq 0.55$, so that $L^* \doteq (0.55/\sqrt{k}|L)$.

6.5 Limitations of Combined Significance Tests.

In spite of the intuitive appeal of using combined test procedures to combine tests of treatment effects, there frequently are problems in the interpretation of results of such a test of the significance of combined results (see e.g., Adcock, 1960; or Wallis, 1942). Just what can be concluded from the results of an omnibus test of the significance of combined results? Recall that the null hypothesis of the combined test procedure is

$$H_0: \theta_1 = \theta_2 = \ldots = \theta_k = 0;$$

that is, H_0 states that the treatment effect is zero in every study. If we reject H_0 using a combined test procedure, we may safely conclude that H_0 is false. However, H_0 is false if at least one of $\theta_1, \ldots, \theta_k$ is different from zero. Therefore, H_0 could be false when $\theta_1 > 0$ and $\theta_2 = \ldots = \theta_k = 0$. It is doubtful if a researcher would regard such a situation as persuasive evidence of the efficacy of a treatment.

The difficulty in the interpretation of omnibus tests of the significance of combined results stems from the nonparametric nature of the tests. Rejection of the combined null hypothesis allows the investigator to conclude only that the omnibus null hypothesis is false. Errors of interpretation usually involve attempts to attach a parametric interpretation to the rejection of H_0. For example, an investigator might incorrectly conclude that because H_0 is rejected, the treatment effects are greater than zero (Adcock, 1960). Alternatively, an investigator might incorrectly conclude that the average treatment effect $\bar{\theta}$ is positive. Neither parametric interpretation of the rejection of H_0 is warranted without additional a priori assumptions.

An additional assumption that is sometimes made is the assumption that there is "no qualitative interaction." This assumption is essentially that if the treatment effect in any study is positive, then no other treatment effect is negative. That is, all of the treatment effects are of the same sign but not necessarily of the same magnitude. While this assumption may seem innocuous, it is important to recognize that it must be made independent of the data and that it may be false. For example, if a drug that actually has a positive effect on a specific disease also has toxic side effects, it might actually increase the death rate among some (e.g., older or sicker) patients. If some studies have more older or sicker patients, it is plausible that they might obtain negative treatment effects while other studies with younger or healthier patients found positive treatment effects.

An important application of omnibus test procedures is to combine the results of dissimilar studies to screen for any treatment effect. For example, combined test procedures can be used to test whether a treatment has an effect on any of a series of different outcome variables. Combined test procedures can even be used to combine the results of related analyses computed using different parameters such as correlation coefficients or effect sizes.

Omnibus tests of the statistical significance of combined results are poorly suited to the task of drawing general conclusions about the magnitude, direction, and consistency of treatment effects across studies. On the other hand, techniques based on combination of estimates of effect magnitude do support inferences about direction, magnitude, and consistency of effects. Therefore, statistical analyses based on effect sizes are preferable for most applications of meta-analysis.

7.0 Combined Estimation.

When all of the studies have similar designs and measure the outcome construct in a similar (but not necessarily identical) manner, the combined estimation approach is probably the preferred method of meta-analysis (Hedges & Olkin, 1985).

It is difficult to discuss the problem of combining the results of studies in complete generality. The purpose and procedures of research studies obviously vary tremendously even within a discipline. It is usually the case, however, that research studies seek to estimate one or more substantively meaningful parameters. The results of a study can therefore often be summarized via an estimate of that parameter and its standard error. An important special case is the situation where studies examine the effect of a "treatment" and the result of the study is an estimate of the "effect" of this treatment measured in some relevant fashion. In this case, the first step in combined estimation is the selection of an index of effect magnitude. Many different indices of effect magnitude have been used in meta-analysis including the raw mean difference between the standardized difference between treatment and control group means (e.g., Smith & Glass, 1977), the observed minus expected frequency of some outcome like death (e.g., Yusof, Peto, Lewis, Collins, & Sleight, 1985), the risk ratio between treatment and control groups (Canner, 1983) or the simple difference between proportions of some outcome in the treatment and control groups (e.g., Devine & Cook, 1983).

Statistical analysis procedure for meta-analysis using any of these indices of effect magnitude are analogous (Elashoff, 1978; Fleiss, 1973; Gilbert, McPeek, & Mosteller, 1977; Hedges, 1983; Hedges & Olkin, 1985; Mantel and Haenzel, 1959; Sheele, 1966). All involve large-sample theory and differ mainly in the details of calculation of standard errors and bias corrections.

Before discussing examples of modeling procedure in combined estimation, it is useful to consider conceptual issues that have implications for that modeling.

7.1 The Nature of Between-Study Variation.

Between-study variation is defined as variability in the study outcome parameters. Three natural conceptualizations of between-study variation treat this variation as totally systematic (e.g., fixed), totally random (e.g., nonsystematic), or mixed (partially systematic and partially nonsystematic). These three conceptualizations give rise to three different types of models for the results of a series of studies. In the fixed-effects conceptualization, the true or population values of the treatment effects in the study are an (unknown) function of study characteristics. By studying the relationship between study characteristics and treatment effects the data analyst tries to deduce stable relationships that explain essentially all of the variability in study results except for that attributable to within-study sampling variability. The evaluation of a particular explanatory models is part of this process.

The random-effects conception arises from a model in which the treatment effects are not functions of known study characteristics. In this model, the true or population values of treatment effects vary randomly from study to study, as if they were sampled from a universe of possible treatment effects (see Hedges & Olkin, 1985). The random effects conceptualization is consistent with Cronbach's (1980) proposal that evaluation studies should consider a model in which each treatment site (or study) is a sample realization from a universe of related treatments. The primary difference between the interpretation of fixed- and random-effects models is that between-study variation in treatment effects is conceived to be unsystematic in random-effects models and consequently explanation of this variance is not possible. Instead the data analyst usually seeks to quantify this variation by estimating a (treatment by studies interaction) "variance component": an index of the variability of population treatment effects across studies.

Mixed models involve a combination of the ideas involved in fixed- and in random-effects models. In these models, some of the variation between treatment effects is fixed (i.e., explainable) and some is random. Consequently, the data Analyst seeks to explain some of the variation between study results and quantify the remainder by estimating a variance component (Raudenbush & Bryk, 1985, DerSimonian & Laird, 1983). Such models have considerable promise as data analytic tools for situations in which it is useful to treat some of the variability between study results as random.

The most important difference in the outcomes produced by the three types of statistical analyses lies in the standard errors that they associate with the overall (combined) estimate of the treatment effect. Fixed-effect analyses incorporate only within-study variability into the estimate of the standard error of the combined treatment effect. Fixed-effects analyses produce the smallest standard error estimates because they are, in fact, conditional on the known and unknown characteristics of the particular studies that have been done. Random-effects analyses include the between-study variance component in estimates of the standard error of the overall (combined) estimate of the treatment effect, and hence produce standard errors between those of fixed- and random-effects analyses.

7.2 Monitoring Models for Between-Study Difference.

Although most statisticians have a great deal of experience which aids their intuition about the specification and robustness of statistical models within studies, few of us have broad experience or accurate intuition about statistical models for between-study variation. Unlike within-study variation, between-study variation is completely uncontrolled by the investigator (reviewer) and even retrospective information about the sampling units (studies) may be difficult to obtain. Consequently modeling assumptions about between-study variation are likely to be wrong, often horrendously wrong. In such an environment, it is very unwise to depend on statistical analysis procedures that are strongly dependent on specification of a particular model for between-study variation.

Research synthesis requires procedures that are either robust against misspecification of between-study models or procedures that allow monitoring of the adequacy of the model for between-study variation. Nonrobust procedures that do not permit (indeed compel) comprehensive monitoring are a recipe for disaster. Two examples may illustrate the point. The first example concerns weighting of the results of different studies when combining estimates of treatment effects. The most efficient estimate of a common treatment effect uses a weight for each study that is inversely proportional to the standard error of the estimated treatment effect. If you really believe the model, you use it regardless of the weights it assigns. A skeptic might believe that no single study should receive too much weight and therefore place an upper bound on the weight any study may attain. This more robust solution (called partial weighting) was suggested by Yates and Cochran (1938) as an alternative to excessive belief in a model. The Techniques of modern robust estimation obviously have a wide range of possible application in meta-analysis, but less robust procedures may be just as useful if their application is carefully monitored.

A second example concerns the use of monolithic data analysis procedures versus data analysis procedures that are easy to monitor in detail. In one sense it is natural to attack the problem of combining information from different studies by utilizing a comprehensive statistical model and performing all aspects of the analysis simultaneously, for example by maximum likelihood estimation. This procedure is elegant and may have certain technical advantages. Yet simultaneous estimation of all aspects of a model may have the disadvantage that it is difficult to discover parts of the model that are not consistent with the data. Moreover, failures of one aspect of the model may affect estimation of other aspects of the model. The alternative of explicitly computing estimates from each study and then combining those estimates using explicit procedures such as generalized least squares is far less elegant. It may also have some technical disadvantages, but it has the advantage that the adequacy of the model is much easier to monitor. Moreover, failures of one aspect of the model are less likely to spill over and create problems in another part of the model. This issue also arises in econometric modeling where the same tradeoffs are recognized between so-called "full information" modeling and "limited information" modeling. The point here is not that simpler methods of combining results are better. It is that models for between-study differences should be tentative and therefore require serious monitoring. Simpler combination procedures are often easier to monitor and therefore have an advantage that might not be obvious. Methods for monitoring models or for producing robust combination procedures are essential in research synthesis.

7.2.2 Realistic Modeling of Between-Study Differences.

Modeling between-study differences is often fraught with complications that are highly idiosyncratic to particular data set under analysis. Research syntheses can only be helpful if they make a serious effort to incorporate these idiosyncrasies into the data analysis model. Some of the sources of these idiosyncrasies have already been mentioned. For example, publication or reporting bias leads to censoring or truncation at the sampling level of studies. Other censoring or truncation effects may exist within studies and both may need to be explicitly modeled. Similarly, missing data that is not missing at random may be important to consider in the model for the data analysis. Finally the possibility of dependence between supposedly independent studies cannot be ignored. If a given team of investigators produces several studies whose results are more alike than those of other investigators, dependencies are introduced which may need to be incorporated into the model specification.

8.0 Some Statistical Methods for Combined Estimation.

This section presents statistical methods that are frequently used in research synthesis. The outline of methods that follows is intended to be generic and therefore it does not incorporate the complexities that might be added in an actual research synthesis.

8.1 Statistical Methods for Fixed Effects Meta-analysis.

Suppose that $T_1,\ldots,T_k$ are independent estimates of effect magnitude form k studies with sample sizes $n_1,\ldots,n_k$ and unknown population effect magnitude parameters $\theta_1, \ldots, \theta_k$. Assuming that the standard error of T_i is a function of θ_i, denote the standard errors of $T_1,\ldots,T_k$ by $S_1(\theta_1), \ldots, S_k(\theta_k)$ and the estimated standard errors by $S_1(T_1), \ldots, S_k(T_k)$. Assume further that each T_i has a normal asymptotic distribution leading to the large-sample normal approximation

$$T_i \sim N(0_i, S^2(T_i)) \tag{7}$$

8.1.1 Estimating the Overall Average Treatment.

One of the first statistical questions that arises is how to estimate the overall average treatment effect when it is believed that $0_1, \ldots, 0_k$ are very similar. One way of combining the estimates is obviously to take the simple average of $T_1,\ldots T_k$. The most precise combination (i.e., the most efficient estimator of 0 when $\theta_1 = \theta_k = \theta$) is a weighted average that takes the standard error $S_1(T_1),\ldots,S_k(T_k)$ into account. This weighted average is

$$T. = \Sigma_{i=1}^{k} \omega_i \; T_i / \Sigma_{i=1}^{k} \omega_i . \tag{8}$$

where $\omega_i = 1/S^2(T_1)$. One slight refinement of (9) is the iterated estimator $T.^{(j)}$ defined by $T.^{(0)} = T.$ and

$$T.^{(j)} = \Sigma_{i=1}^{k} \omega_i^{(j-1)} . T_i / \Sigma_{i=1}^{k} \omega_i^{(j-1)} . j=1,2 \ldots, \tag{9}$$

where $\omega_i^{(j)} = 1/S_i^2(T.^{(j-1)})$. When $N = \Sigma_{i=1}^{k} n_1$ with n_1/N fixed

(that is, if each study has a large sample size) and if each $S_i(T)$ is a continuous function of T, the estimators T and $T^{(j)}$ have (the same) asymptotic distribution leading to the large-sample normal approximation

$$T \sim N(\theta, S.^2(T.)), \tag{10}$$

where

$$S.^{-2}(T.) = \Sigma_{i=1}^{k} S_i^{-2} (T.) . \tag{11}$$

This result can be used to compute tests of significance and confidence intervals for based on T.. For example a 100 (1 − α) percent confidence interval for θ is given by

$$T. - z_{\alpha/2} S.(T.) \le \theta \le T. + z_{\alpha/2} S.(T.), \tag{12}$$

where $z_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution. Alternatively, a test of the hypothesis that $\theta = 0$ uses the statistic

$$Z = T. / S.(T.), \tag{13}$$

which is compared to the critical values of the standard normal distribution. If the T_i, $i=1,\ldots,k$, are symptotically efficient (have asymptotic variances equal to the Cramer-Rao lower bound), then T. is asymptotically efficient.

Note that weighted combinations of estimators using estimated weights and their iterated counterparts have identical large sample properties. Thus the decision about whether to iterate to obtain the estimates of θ used in the weights depends on the form of $S_i^2(\theta)$ and on the small sample properties of T.. If $s_i^2(\theta)$ is almost independent of θ for the plausible values of θ in the problem at hand, iteration is unlikely to have much effect. On the other hand, if $S_i^2(\theta)$ changes considerably as a function of plausible values of θ, iteration may change the value of the estimate by a considerable amount. Note also that even if the T_i are unbiased estimates of θ, T. is usually biased. The iterated estimators $T.^{(j)}$ will usually tend to be less biased than T.

Another point is that the regularity condition that the n_i/N remain fixed as $N \to \infty$. If we let N increase by increasing k and letting the n_i/N 0, then a variety of problems arise. For example, under this condition T. is not even consistent if the T_i are biased estimators of θ (see Neyman & Scott, 1948).

8.1.2 Testing Homogeneity of Treatment Effects.

Combining estimates of effect magnitude across studies is reasonable if the studies have a common population effect magnitude θ. In this case, properties of T_2 and on the dependence of the variance of the T_i on θ_i. If the variance $S_i^2(\theta)$ is almost independent of 0, then the iteration process will not change the estimates very much. If $S^2(0)$ is greatly influenced by θ, then iteration could change the estimate of β considerably. The iterated estimators are also likely to be less biased than the uniterated estimators.

This result is often useful in situations where the model for the data is well understood, but the obvious means of obtaining estimates, such as the method of maximum likelihood, require complicated iterative methods. The present method yields estimators that are simple to compute but have the same large-sample properties as maximum-likelihood estimators. In addition the diagnostic procedures routinely used in weighted least squares can be applied in the usual manner.

If the investigator proposes a linear model for θ, it may be desirable to see if the model is reasonably consistent with the estimates $T_1,\ldots,T_k$. If the model does not seem reasonably consistent with the data, then the entire analysis should be suspect. Iterated estimators, in particular, would not be expected to perform well if the model is misspecified. Whenever $k > p$, a natural test of model specification arises in connection with the estimator presented above. The test given below provides a way to check that the data are consistent with the proposed model.

If $\beta = 0$, then the statistic

$$H_M = T'V^{-1}(T)X(X'V^{-1}(T)X)^{-1}X'V^{-1}(T)T \tag{20}$$

is distributed approximately as a chi-square with p degrees of freedom. The statistic H_M can be used as for a simultaneous test that $\beta_1 = \ldots \beta_p = 0$, or alternatively as a simultaneous test that $\theta_1 = \ldots \theta_k = 0$. More significantly H_M is used in the calculation of a test for goodness of fit or specification of the linear model. If $k > p$ and the model $\theta = X\beta$ is correctly specified, then the statistic

$$H_E = T'V^{-1}(T)T - H_M \tag{21}$$

has an approximate distribution given by

$$H_E \sim \chi^2_{k-p}$$

The statistic H_M is the (weighted) sum of squares due to the regression model and the statistic H_E is the (weighted) sum of squares about the regression plane (the r sum of squares). Thus the test for model misspecification is a test for larger than expected residual variation. If H_E is large or significant, the investigator might use any of the standard tools of regression analysis (e.g., the examination of residuals, the search for influential observations, etc.) to look for problems with the model.

Note that the linear model analyses described above can all be performed with standard packaged computer programs such as SAS PROC GLM. The weighted regression (or analysis of variance) is performed by simply specifying the weight for each case (each T_i) as

$$\omega_i = 1/s_i^2(T_i).$$

The regression coefficients are printed directly and the variances of the estimated regression coefficients are the diagonal elements of the inverse of the (X'WX) matrix. The statistics H_M and H_E are printed as the weighted sum of squares due to the regression model and the weighted error sum of squares about the regression plane.

8.2 Statistical Methods for Random-Effects Meta-analysis.

Again suppose that $T_1,\ldots,T_k$ are independent estimates of treatment effects from k experiments with (unknown) population treatment effects . Again denote the standard error of T_i given by $S_i(T_i)$. Assume as before that the T_i are approximately normally distributed. Now, however, introduce the random-effects model that , are sampled from a hyperpopulation of treatment effects. Often the are assumed to be normally distributed. The object of the analysis is to estimate the mean and variance (the hyperparameters) of the distribution of population treatment effects, and to test the hypothesis that $= 0$.

A distribution-free approach to estimating is analogous to the procedure used to estimate the variance component in the one-factor random effects analysis of variance. The estimate is given by

$$\hat{\sigma}^2_\theta = s_T^2 - \Sigma_{i=1}^k s_i^2(T_i)/k, \tag{22}$$

where S_T^2 is the usual sample variance of $T_{1,2} \ldots, T_k$ (see Hedges and Olkin, 1985). More complex procedures for estimating σ_θ^2 under various distributional assumptions on the θ_i are given in Champney (1983), Raudenbush and Bryk (1985), and Hedges and Olkin (1985).

The usual estimate of $\bar{\theta}$ is the weighted mean

$$T_{\cdot}^{*} = \sum_{i=1}^{k} \omega_i^{*} T_i / \sum_{i=1}^{k} \omega_i^{*}, \tag{23}$$

where $\omega_i^{*} = 1/[\hat{\sigma}_\theta^2 + S_i^2(T_i)]$.

The weighted mean $T^{*}.$ is approximately distributed

$$T_{\cdot}^{*} \sim N(\bar{\theta}, \sigma_{*}^2), \tag{24}$$

where $\sigma_{*}^{-2} = \sum_{i=1}^{k} \omega_i^{*}$.

Consequently, an approximate confidence interval for $\bar{\theta}$ is given by

$$T_{\cdot}^{*} - z_{\alpha/2}\sigma_{*} \leq \theta \leq T_{\cdot}^{*} + z_{\alpha/2}\sigma_{*} \tag{25}$$

where $z_{\alpha/2}$ is defined as in (12). Note that the weights ω_i^{*} used in (24) are not the same as the weights ω_i used in (9) unless σ_θ^2 is exactly zero. Usually σ_θ^2 is larger than zero and consequently $T^{*}.$ differs from T.. Moreover, the standard error σ_{*} of $T^{*}.$ is usually larger (often much larger) than the standard error σ of T.. As a result, overall treatment effects that are significantly different from zero in a fixed-effects analysis may not be significant in a random-effects analysis. The difference, of course, results from differences in the conceptualization of the model, and in what counts as sampling error.

8.3 Statistical Methods for Mixed-Effects Meta-analysis.

Statistical methods for mixed-effects meta-analyses have received less complete treatment in the literature than have fixed- and random-effects models. There is a great deal of work in progress on meta-analysis with mixed-effects models. This work shows potential for resolving differences between fixed- and random-effects approaches. One useful treatment of mixed-effects meta-analysis in the context of educational research is Raudenbush and Bryk (1985).

8.4 Limitations of Statistical Methods

One of the most important issues faced by any reviewer is that of the limitations of statistical methods for answering the types of questions that are addressed by the review. In spite of a large number of apparently relevant research studies, it is often the case that the actual amount of specifically relevant empirical information is quite small. For example, in some areas the large bulk of the research studies fail to meet methodological standards that are now considered essential. In other cases the reviewer's interest is focused on very specific situations which are not adequately represented or differentiated in the extant research literature. In still other situations, individual studies (such as very small clinical trials) are so inadequate that even pooling a fairly large number of such studies still leads to parameter estimates with very large standard errors.

9.0 The Place of Reviews in Scientific Explanation.

The function of research reviews in science is to do more than collect and tabulate research results. It is no accident that the most prestigious journals that publish research reviews seek reviews that are "critical," "integrative," or "synthetic" or that the words "research synthesis" are sometimes used a a synonym for "research reviewing." These descriptions of research reviews imply something beyond tabulation, they suggest that reviews will not only present empirical results and generalizations but will offer explanations in support of those generalizations. The important issue is that the explanation consists of more than the empirical generalization and a sketchy summary of the data on which it is based. All explanations relate the phenomenon to be explained to other ideas that are presumably understood and perceived to be relevant by the recipient of the explanation. It usually involves demonstrating the ways in which the new phenomenon fits into patterns that are familiar and are perceived to be relevant in some way. We understand by discovering the many linkages between a new phenomenon and existing beliefs, relationships, and empirical data. The function of the explanation is to make the linkages clear, to make obvious the ways in which the new phenomenon fits into the matrix of background beliefs, theories, empirical data, and relationships. Note that a perfectly correct generalization alone is not necessarily a good explanation if the links to appropriate context are not made obvious.

Meta-analysis has prompted a considerable amount of debate in some quarters during the last ten years. Perhaps the controversy is understandable as a natural consequence of the introduction of a new research paradigm. Some of the controversy seems to be based on a misunderstanding of meta-analysis. Part of it may be an inevitable consequence of the rough edges of the first few studies produced by any new research paradigm. But some of the criticisms seem to suggest that meta-analyses have sometimes failed as explanations. This is disturbing because research reviews using meta-analysis must succeed as explanations if they are to be useful. Moreover, if they are to have any lasting impact they must be convincing explanations to the researchers and policy makers who are most knowledgeable about the subject matter under review. Meta-analyses that convince only other experts in meta-analysis are not useful. A crucial question is why do explanations based on meta-analyses sometimes fail as explanation and how can they be improved.

9.1 Why Meta-analyses Face Difficulty as Explanation.

Meta-analyses and indeed any research reviews that offer precise conclusions are likely to face some difficulties as credible explanation. The reason is that to make generalizations across studies the reviewer must ignore a great many differences among those studies. Glass, McGaw, and Smith (1981) have argued that generalizing across studies is not logically different from generalizing across individual subjects within an experiment. While this may be true at some level, the analogy ignores an important perceptual difference between generalizations within and across studies. The experimental paradigm predisposes researchers to view the generalization across subjects as natural because, by definition, the differences between subjects are "experimental errors" (see Cronbach, 1957). The few systematic individual difference variables recognized in the experimental paradigm are incorporated into the design and all other differences are by definition nonsystematic.

Differences between studies, on the other hand, are viewed as systematic because the same experimental paradigm stresses the importance of the design of research studies. A great deal of the training and professional effort of researchers is devoted to learning about, planning, and implementing systematic aspects of research studies that make one study different from another. I emphasize that the differences between studies are viewed by knowledgeable researchers as systematic, because researchers strive to make their studies systematically different from those of other researchers to obtain new information. They do so because they believe these differences could have an effect on the results of the study. Moreover, the differences between studies are not usually unidimensional. The design and execution of research studies is so complex that even "similar" studies often differ in many ways. For this reason, it would be expected that researchers would have difficulty with any method of generalizing across studies, because such methods implicitly relegate the many complex and important differences between studies to the status of "error" or unsystematic variation. Researchers are likely to find even more problematic, statistical methods for generalizing across studies that explicitly define all of the unmodeled variation between study results to be sampling error. Conventional statistical methods (such as t tests, analysis of variance and multiple regression analysis) applied to effect sizes are examples of this type. Statistical methods developed specifically for meta-analysis separate variation between studies that is due to sampling error within studies from that which is due to systematic variation between studies.

The most persistent criticisms of meta-analysis stem, in part, from the perspective of researchers who feel that differences among studies and among their results are systematic and that meta-analysis fails in some way to recognize those differences. Perhaps the most consistent of these criticisms of meta-analysis (which could be criticisms of any review) have come to be called the "apples and oranges" criticism and the "garbage-in garbage-out" criticism.

The apples and oranges criticism maintains that meta-analysis combines evidence from studies which do not have the "same" procedures, independent variables, or dependent variables. Thus meta-analysis is combining the incommensurable because studies exhibit systematic differences. Another statement of essentially the same criticism (Presby, 1978) is that combining research studies into overly broad categories obscures important differences between those studies and their results. In each case, the fundamental issue is the breadth of constructs that are the "same," and the critic's position is that only aggregation across a rather narrow range of treatment, control, and outcome constructs is sensible.

The "garbage-in garbage-out" criticism (Eysenck, 1978) is that by abandoning "critical judgment" about the quality of research studies reviewed, meta-analysis placed too much emphasis on studies of low quality. Because studies of low quality are presumably subject to many biases, they cannot be the foundation of reliable knowledge. Meta-analysis therefore becomes another case of garbage-in garbage-out. Although the criticism concerns the question of methodological quality, it is firmly rooted in the conception that there are systematic differences (in methodology) between studies that influence study results.

9.2 Improving Meta-analysis in the Service Scientific Explanation.

The improvement of meta-analysis as explanation depends on greater attention to both methodological detail and to persistent criticisms of meta-analysis. Critics tell us why they do not find meta-analyses to be convincing as explanation. Attempts to respond to those criticisms (where they do not conflict with other requirements of methodology) are likely to yield more persuasive meta-analyses. Many of these criticisms are among the issues discussed in earlier sections of this chapter, but two general issues emerge. One is the issue of specificity versus generality of constructs. The other is the appropriate use of quantitative methods.

The issues of specificity arise because researchers tend to think of studies in terms of specific and rather narrow constructs. This tendency toward specificity is reflected in the usually narrow choice of constructs in conventional reviews (Cook & Leviton, 1980). Meta-analyses are likely to be more credible as explanation if they use (or at least distinguish) constructs of treatment, control, and outcome that are relatively narrow and relatively specific to the research domain at hand. Meta-analyses are also likely to be more credible if they use conceptions of study quality that recognize the specific difficulties associated with the domain under study. By treating between-study differences in rather specific ways, meta-analyses will offer a richer variety of connections with researcher's conceptualizations of the research domain.

The issues of appropriate use of quantitative methods might be interpreted to include all issues of the formal (mathematical) appropriateness of statistical methods in a given situation. More important is the question of when should statistical methods be used given that they are formally correct. Researchers are not always comfortable with the use of quantitative method to empirically "define" the differences among studies that deserve consideration. For example, the argument that study quality can be defined empirically by determining which groups of studies give different answers has not always been persuasive. Critics seem to be saying that quantitative analyses cannot carry the whole load. Meta-analyses are likely to be more persuasive if they use qualitative methods to determine interesting differences among studies. Researchers know both that quantitative methods cannot resolve all questions and that these methods must be guided and set in context by qualitative analysis. Qualitative information that is not explicitly coded as between-study differences has an important role in interpretation and should not be neglected even if it requires rather lengthy descriptions of important aspects of individual studies (Light & Pillemer, 1984).

The net effect of these suggestions would be to make meta-analyses look more like conventional narrative reviews, involving perhaps fewer studies distinguishing narrower constructs, and providing more detailed qualitative and conceptual arguments. In fact, earlier conventional reviews of an area may be a model for conceptualization and level of operational detail that are appropriate. The most persuasive meta-analysis is likely to be one that combines the strengths of qualitative reviews and those of serious quantitative methodology.

References

Adcock, C. J. (1960). A note on combining probabilities. Psychometrika, 25, 303-305.

Bakan, D. (1966). The test of significance in psychological research, Psychological Bulletin, 66, 423-437.

Barnett, V., & Lewis, T. (1978). Outliers in Statistical Data. New York: John Wiley.

Bozarth, H. D., & Roberts, Jr., R. R. (1972). Signifying significant significance. American Psychologist, 27, 774-775.

Birge, R. T. (1932). The calculation of errors by the method of least squares. Physical Review, 16, 1-32.

Bracht, G., & Glass, G. V. (1968). The external validity of experiments. American Educational Research Journal, 5, 437-474.

Campbell, D. T. (1969). Definitional versus multiple operationalism. et. al, 2, 14-17.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and Quasiexperimental Designs for Research. Chicago: Rand McNally.

Canner, P. L. (1983). Aspirin in coronary heart disease: A comparison of six clinical trials. Israel Journal of Medical Sciences, 19, 413-423.

Chalmers, T. C. (1982). The randomized controlled trial as a basis for therapeutic decisions. In J. M. Lachin, N. Tygstrup, & E. Juhl (Eds.). The Randomized Clinical Trial and Therapeutic Decisions. New York: Marcel Dekker.

Champney, T. F. (1983). Adjustments for Selection: Publication Bias in Quantitative Research Synthesis. Unpublished doctoral dissertation. The University of Chicago.

Clarke, F. W. (1920). A redetermination of atomic weights. Memoirs of the National Academy of Science, 16(3), 1-48.

Cochran, W. C. (1937). Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society (Supplement), 4, 102-118.

Cochran, W. C. (1943). The comparison of different scales of measurement for experimental results. Annals of Mathematical Statistics, 14, 205-216.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation. Chicago: Rand McNally.

Cook, T. D., & Leviton, L. C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. Journal of Personality, 48, 449-472.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. Journal of Personality and Social Psychology, 37, 131-146.

Cooper, H. M. (1984). The Integrative Research Review: A Systematic Approach. Beverly Hills: Sage Publications.

Cooper, H. M. & Arkin, R. M. (1981). On quantitative reviewing. Journal of Personality, 49, 225-230.

Crain, R. L. & Mahard, R. E. (1983). The effect of research methodology on desegregation-achievement studies: A meta-analysis. American Journal of Sociology, 88, 839-684.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. American Psychologist, 12, 671-684.

Cronbach, L. J. (1980). Toward Reform of Program Evaluation. San Francisco: Jossey-Bass.

DerSimonian, R., & Laird, N. (1983). Evaluating the effectiveness of coaching for SAT exams: A meta-analysis. Harvard Educational Review, 53, 1-15.

Eagley, A. H. & Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinents of sex differences in influenceability: A meta-analysis of social influence studies. Psychological Bulletin, 90, 1-20.

Eagley, A. H. & Crowley, M. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. Psychological Bulletin, 99.

Elashoff, J. D. (1978). Combining the results of clinical trials. Gastroenterology, 28, 1170-1172.

Eysenck, H. J. (1978). An exercise in mega-silliness. American Psychologist, 33, 517.

Fisher, R. A. (1932). Statistical Methods for Research Workers (4th ed.) London: Oliver & Boyd.

Fleiss, J. L. (1973). Statistical Methods for Rates and Proportions. New York: John Wiley.

George, E. O. (1977). Combining Independent One-sided and Two-sided Statistical Tests — Some Theory and Applications. Unpublished doctoral dissertation, University of Rochester.

Giaconia, R. M. & Hedges, L. V. (1982). Identifying features of effective open education. Review of Educational Research, 52, 579-602.

Gilbert, J. P., McPeek, B., & Mosteller, F. (1977). Progress in surgery and anesthesia: Benefits and risks of innovation therapy. In J. Bunker, B. Barnes, and F. Mosteller (Eds.). Costs, Risks, and Benefits of Surgery. New York: Oxford University Press.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Glass, G. V. & Smith, M. L. (1979). Meta-analysis of the relationship between class size and achievement. Educational Evaluation and Policy Analysis, 1, 2-16.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in Social Research. Beverly Hills: Sage Publications.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82, 1-20.

Hawkins, D. M. (1980). Identification of Outliers. London: Chapman Hall.

Hedges, L. V. (1983). Combining independent estimators in research synthesis. The British Journal of Mathematical and Statistical Psychology, 36, 123-131.

Hedges, L. V. (1984). Estimation of effect size under normal nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. Journal of Educational Statistics, 9, 61-85.

Hedges, L. V. & Olkin, I. (1980). Vote counting methods in research synthesis. Psychological Bulletin, 88, 359-369.

Hedges, L. V. & Olkin, I. (1983). Clustering estimates of effect magnitude from independent studies. Psychological Bulletin, 93, 563-573.

Hedges, L. V. & Olkin, I. (1985). Statistical Methods for Meta-analysis. New York: Academic Press.

Hunter, J. E., Schmidt, F. L., & Jackson, J. B. (1982). Meta-analysis: Cumulating findings across research. Beverly Hills: Sage.

Lane, D. M. & Dunlap, W. P. (1978). Estimating effect sizes: Bias resulting from the significance criterion in editorial decisions. British Journal of Mathematical and Statistical Psychology, 31, 107-112.

Light, R. J., & Pillemer, D. B. (1984). Summing Up: The Science of Reviewing Research. Cambridge, Massachusetts: Harvard University Press.

Linn, M. C., & Peterson, A. C. (1985). Emergence and characterization of sex differences in spatial ability. Child Development, 56, 1479-1498.

Madow, W. G., Niesselon, H., & Olkin, I. (1983). Incomplete data in sample surveys: Vol. I, Report and case studies. New York: Academic Press.

Madow, W. G., & Olkin, I. (1983). Incomplete data in sample surveys: Vol. 3. Proceedings and the symposium. New York: Academic Press.

Madow, W. G., Olkin, I., & Rubin, D. B. (1983). Incomplete data in sample surveys: Vol. 2. Theories and bibliographies. New York: Academic Press.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. Journal of the National Cancer Institute, 22, 719-748.

Melton, A. W. (1962). Editorial. Journal of Experimental Psychology, 64, 553-557.

Miller, R. G. (1981). Simultaneous Statistical Inference. (2nd Ed.). New York: Springer-Verlag.

Mudholkar, G. S. & George, E. O. (1979). The logit method for combining probabilities. In J. Rustagi (Ed.). Symposium on Optimizing Methods in Statistics (pp. 345-366). New York: Academic Press.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. Econometrica, 16, 1-32.

Orwin, R. G., & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. Psychological Bulletin, 97, 134-147.

Pearson, K. (1933). On a method of determining whether a sample of given size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. Biometrika, 25, 379-410.

Presby, S. (1978). Overly broad categories obscure important differences American Psychologist, 33, 514-515.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. Journal of Educational Statistics, 10, 75-98.

Rocke, D. M., Downs, G. W., & Rocke, A. J. (1982). Are robust estimators really necessary? Technometrics, 24, 95-101.

Rosenfeld, A. H. (1975). The particle data group: Growth and operative. Annual Review of Nuclear Science, 555-599.

Rosenthal, R. (1984). Meta-analytic Procedures for Social Research. Beverly Hills: Sage Publications.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63, 581-592.

Sheele, P. R. (1966). Combination of log-relative risks in retrospective studies of disease. American Journal of Public Health, 56, 1745-1750.

Simpson, E. H. (1954). The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, Series B, 13, 238-241.

Smith, M. L. (1980a). Publication bias in meta-analysis. Evaluation in Education: An International Review Series, 4, 22-24.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.

Stampfer, M. J., Goldhaber, S. Z., Yusuf, S., Peto, R., & Hennekins, C. H. (1982). Effects of intraveneous streptokinase on acute myocardial infarction. New England Journal of Medicine, 307, 1180-1182.

Sterling, T. D. (1959). Publications decisions and their possible effects on inferences drawn from tests of significance—or vice versa. Journal of the American Statistical Association, 54, 30-34.

Stigler, S. M. (1977). Do robust estimators work with real data? Annals of Statistics, 5, 1055-1098.

Stock, W. A., Okun, M. A., Haring, M. J., Miller, W., Kinney, C., & Cuervost, R. W. (1982). Rigor in data synthesis: A case study of reliability in meta-analysis. Educational Researcher, 11, 10-14, 20.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, Jr., R. M., (1949). The American Soldier, Volume 1. Adjustment During Army Life. Princeton: Princeton University Press.

Thomas, J. R. & French, K. E. (1985). Gebder differences across age in motor performance: A meta-analysis. Psychological Bulletin, 98, 260-282.

Tippett, L. H. C. (1931). The Method of Statistics. London: Williams and Norgate.

Tukey, J. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley.

Wallis, W. A. (1942). Compounding probabilities from independent significance tests. Econometrica, 10, 229-248.

Webb, E., Campbell, D., Schwartz, R., Sechrest, L., & Grove, J. (1981). Unobstrusive measures: Nonreactive research in the social sciences. Boston: Houghton Mifflin.

Woolf, B. (1955). On estimating the relation between blood group and disease. Annals of Human Genetics, 19, 251-253.

Wortman, P. M. (1981). Randomized clinical trials. In P. M. Wortman (Ed.). Methods for Evaluating Health Services. Beverly Hills: Sage Publications.

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. Journal of Agricultural Research, 28, 556-580.

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. Progress in Cardiovascular Diseases, 27, 335-371.

DISCUSSION

Chao W. Chen,
U.S. Environmental Protection Agency

Dr. Hedges has given an excellent discussion about general issues that one should consider when combining information from different studies. However, the methodologies proposed in his presentation do not appear to have much usefulness for most problems that the U.S. Environmental Protection Agency (EPA) encounters in environ- mental management. A major difficulty in environmental management is that scientists often cannot provide precise estimates of environmental risk posed by a pollutant because of a lack of scientific knowledge and data. As is usually the case, very little is known about the potential risks of chemicals in the air, water, or workplace. Our knowledge is even weaker with respect to the mechanism of carcinogens. Yet EPA is frequently called upon to make decisions on the management of environmental risk, in the face of such an enormous uncertainty. A simple example is about the issue of whether one should combine benign and malignant tumors in statistical evaluations of carcinogen data. This decision could make a big difference in classifying an agent as to whether or not it is carcinogenic. For instance, suppose that three benign and four malignant neoplasms of the same cell type are found in a group of 50 animals, versus none in a control group of 50 animals. When these incidence data are analyzed separately, none of them is statistically significant (one-sided, $p > 0.05$) with the use of the Fisher Exact Test. However, when benign and malignant tumors are combined (0/50 vs. 7/50), the incidence is highly significant (one-sided, $p < 0.007$). Since there are scientific reasons for favoring and opposing combination of neoplasms, it would be more appropriate to reflect this uncertainty in risk assessment and management. The "meta-analysis" which considers only statistical variability is not capable of taking into account this dynamic nature of the problem.

The methodologies presented by Hedges consist of two parts: namely, combining significance tests and statistical analysis of combined effect size. The procedure for analyzing combined effect size is mainly an analysis-of-variance (ANOVA)-type approach, which may not be adequate for the kind of problem usually encountered by the EPA in the area of risk assessment. Some of the procedures used to combine significance tests are multiple comparison tests in nature (e.g., minimum p method), which is certainly not the objective of combining information from different studies. In the minimum p method, the null hypothesis $H_0 : H_{01} = H_{02} = \ldots = H_{0k}$ is rejected if

$$p_m < 1 - (1 - \alpha)^{1/k}$$

where p_m is the minimum of $p_1, p_2, \ldots, p_k$. This procedure is simply a simultaneous inference test for a family of k independent statements (S_i, $i = 1, 2, \ldots, k$) with a family error rate α. This can be easily seen as follows:

$$p_i = \Pr(S_i \text{ is incorrect})$$

$$1-p_i = \Pr(S_i \text{ is correct})$$

$$1-\alpha = \Pr(\text{all } S_i \text{ are correct}) = \prod_i (1-p_i) < (1-p_m)^k$$

Therefore, the null hypothesis H_0 is rejected if

$$p_m < 1 - (1-\alpha)^{1/k}$$

Similarly, the use of Fisher's procedure of combining information from different studies may not be appropriate, as the following example demonstrates. Suppose two studies are identical except for the sample size and are summarized in 2x2 tables as follows:

14	20
10	10

140	200
100	100

The p values for these two studies are, respectively, 0.60 and 0.04, using the one-sided Fisher Exact Test. The combined result by Fisher's procedure is $-2\Sigma_i \log(p_i) = 3.2$, which has a one-sided p-value of 0.52 under H_0. Clearly, it is not appropriate to perform such a statistical procedure when one of the two null hypotheses is already rejected. This example demonstrates that combining significance tests as proposed may not be meaningful.

The last example I will present is the problem of combining cancer risk assessment results. Since there are many uncertainties associated with each step in risk assessment, the problem of combining these results is complex and obviously could not be resolved by ANOVA-type statistical approaches. Table 1 provides hypothetical information on a suspect carcinogen which induced liver tumors in both male and female B6C3F1 mice via either the gavage or inhalation routes of exposure. This suspect carcinogen also induced leukemia in male and female F344 rats by inhalation, but failed to induce tumors in Osborne-Mendel rats in a lifetime gavage bioassay or in Sprague-Dawley rats in a one-year inhalation study. There is a debate in the scientific community with regard to the significance of B6C3F1 mouse liver tumors to humans because of the high spontaneous tumor rates in these animals. The problem facing EPA is how to use this and other information to arrive at a conclusion or decision as to whether this suspect carcinogen could cause cancer in humans, and if it is a human carcinogen, to determine its risk to humans.

Assuming that the compound is a human carcinogen, the risk estimates (risk at 1 μ g/m^3) are calculated with a dose-response model that is linear at low doses. It is generally accepted that low-dose linearity provides a plausible upper-bound estimate of risk. However, some fragmental information indicates that the true shape of the dose-response curve may be sublinear, but the degree of curvature is not known. In combining statistical significance tests or effect sizes (risk estimates), it is desirable to take into account all the available scientific information, which is itself very uncertain. Statistical procedures, such as meta-analysis, that take into account only the sampling variability, are clearly not adequate.

TABLE 1. SIGNIFICANCE LEVELS, p, OF TREND TEST FOR VARIOUS BIOASSAYS AND ESTIMATES OF CANCER RISK AT 1 μg/m^3, CALCULATED ON THE BASIS OF THESE STUDIES

				p-values			Risk
Animals	Sex	Route of exposure	Site	Malignant	Benign	Combined	estimates at 1 μg/m^{3a}
B6C3F1 mice	M	Gavage	Liver	0.018	N.S.	N.S.	6 x 10^{-5}
	F	Gavage	Liver	0.003	N.S.	N.S.	7 x 10^{-6}
	M	Inhalation	Liver	0.001	0.001	0.001	5 x 10^{-7}
	F	Inhalation	Liver	0.001	N.S.	0.001	4 x 10^{-7}
Osborne-Mendel rats	M	Gavage	No response				
	F	Gavage	No response				
F344 rats	M	Inhalation	Leukemia	0.004			6 x 10^{-7}
	F	Inhalation	Leukemia	0.050			1 x 10^{-7}
Sprague-Dawley rats	M	Inhalation	No response				
	F	Inhalation	No response				

[a] These estimates are calculated on the basis of malignant tumors alone. In practical risk assessment, estimates could also be calculated on the basis of benign and/or benign and malignant tumors combined. For ease of presentation, they are not presented here.

N.S. = Not significant (p > 0.05).

DISCUSSION

James M. Landwehr, AT&T Bell Laboratories

Combining the results from several studies through performing a meta-analysis of them is clearly becoming both important and widely practiced, especially in the social sciences. Prof. Hedges has, in this paper as well as in other publications, carefully and systematically presented the statistical methodology of meta-analysis. In this discussion I will briefly give an overall framework for statistical applications that I find useful. Then I will relate meta-analysis to this framework, identify parts of meta-analysis that need further attention, and make several suggestions concerning the methodology.

Before proceeding, let me briefly state my main point and general conclusion. Meta-analysis should not be thought of as some completely new and different kind of statistical methodology. Its steps fit nicely within the framework that we use for statistical applications. Doing a meta-analysis well, however, requires much care with the underlying assumptions and the conclusions.

I like to think of a statistical application as having five main parts. The first is problem formulation, including the design of the study or experiment, and data collection. The second step can be thought of as data analysis, or descriptive statistics, or exploratory data analysis; this step often uses graphical displays extensively. Following the data analysis is construction of more formal models, which can either be deterministic and/or stochastic. The fourth step involves formal statistical inference, perhaps expressed in terms of parameters of a model constructed previously, along with diagnostic checking of the model. Finally, the results must be presented in a way that is informative to the audience.

Clearly there is some subjectivity in the definitions of these steps and overlap between them, but it is not necessary to be terribly precise. I will relate some of the terms, methods, and issues of meta-analysis to these steps.

Problem Formulation, Study Design, Data Collection. Some of the important points of meta-analysis related to this step are the following: the breadth of questions asked in the separate studies and in the meta-analysis; having clear definitions of the variables; making sure that it is reasonable to treat the parameters as being the same across studies; and considering whether or not there is a "representative sample" of studies, subjects, and treatments. Judgments of the quality of the individual studies are an important part of the data evaluation process. Prof. Hedges discussed these issues carefully and extensively.

Data Analysis, Descriptive Statistics, Graphical Displays. This is an important part of statistical applications that deserves more emphasis in meta-analysis. Plotting the data in various ways and studying the plots should be done at an early stage of any application. This helps the analyst to become familiar with the data, to find possible errors, to generate new ideas and hypotheses, and to get initial views on whether or not the data will answer questions of interest. Sometimes simple plots do give clear and convincing answers to the important questions, so this is all the analysis that is really necessary.

Suppose we have an estimated parameter T_i with its estimated standard deviation $\hat{s}_i(T_i)$ from the i^{th} study, for i from 1 to K. We should display these values somehow. For example, construct a plot with estimated parameter value on the ordinate and study index on the abscissa. The ordinate has a * at T_i and a vertical line from $T_i - 2\hat{s}_i(T_i)$ to $T_i + 2\hat{s}_i(T_i)$, and the abscissa is the index i. Examining such a plot gives a rough idea of whether or not the differences among the estimates T_i are consonant with the internal confidence intervals from the individual studies. That is, does it appear that the studies are consistently estimating a common parameter or not? The plot might indicate particular studies that seem quite different from the rest, either in the estimate T_i or its estimated variability.

For any moderator variable (i.e., explanatory variable) X_i that measures some relevant characteristic of the i^{th} study, we should also construct the corresponding plot in which the abscissa is X_i rather than the study index i. For the ordinate plot the same vertical line as before. This plot shows the general relationship between the parameter being studied and the moderator variable X. Many such plots should be constructed and studied using whatever variables are available for analysis.

In their book, Hedges and Olkin (1985) do give an example of a plot of the type described two paragraphs earlier. However, it appears on page 252 of the book! It is the first plot of data that is shown and follows much discussion of statistical models and tests. I suggest that such plots should be constructed and used at the very beginning of any meta-analysis, before getting into more complicated modeling, testing, and estimation.

Models, Deterministic and/or Stochastic. The canonical model in meta-analysis seems to be the following. From the i^{th} study we have some parameter θ_i, its estimate T_i from sample size n_i, and a formula for the standard deviation of T_i, which is denoted $s_i(\theta_i)$. Generally we assume that T_i is approximately normally distributed. From each study there is essentially one data point, T_i. The key difference between this situation and most other statistical problems is that here we also have an estimated standard deviation, namely $s_i(T_i)$, for this data point that is calculated without using any information from other data points. Typically in statistical problems we must use the variability across data points in order to calculate a standard deviation that applies to a particular data point.

Here is a specific model illustrating these concepts taken from Chapter 5 of Hedges and Olkin (1985). Suppose the i^{th} study has experimental and control groups and response variable Y. Assume that in the experimental group the observation on the j^{th} subject, $Y_j^{E_i}$, is distributed normally with mean μ^{E_i} and variance σ_i^2, and in the control group $Y_j^{C_i}$ is distributed normally with mean μ^{C_i} and variance σ_i^2. Define the parameter for the meta-analysis to be $\theta_i=(\mu^{E_i}-\mu^{C_i})/\sigma_i$; this standardized mean difference is called the *effect size* for the i^{th} experiment. It is invariant under linear transformation of the response variable for the study. In meta-analysis there must be reason to believe that the population effect sizes θ_i are measuring the same conceptual and operational quantity across different studies. The value θ_i can be estimated by $T_i=(\bar{Y}^{E_i} - \bar{Y}^{C_i})/s_i$, where s_i is the pooled within-group sample standard deviation. Then $\sqrt{n_i}\,T_i$ has a noncentral t–distribution with noncentrality parameter $\sqrt{n_i}\,\theta_i$ where $n_i = n_i^E n_i^C/(n_i^E + n_i^C)$. This implies that the expected value of T_i is approximately θ_i and its variance is approximately $1/n_i$.

Let's consider a few of the assumptions underlying this model, apart from the obvious one that the θ_i's may have a common value θ across studies. We assume that the variables Y in the different studies are measuring the same underlying theoretical construct. We also assume that the variables in the different studies are linearly equivalent; that is, that they are really measuring the same operational parameter apart from separate linear transformations in each study. Another assumption is that this parameter θ is meaningful. We assume that in each study there is equal variance within the control and treatment groups, and that the distributions are normal. We also assume independence across studies and that it is reasonable to summarize the responses from the i^{th} study by the single variable T_i.

Suppose the observed values T_i and $s_i(T_t)$ imply, using the statistical model, that the θ_i's are not all equal to a common value θ. This type of conclusion is an important output of a meta-analysis and could be substantively important. An alternate conclusion, however, might be that one or more of the assumptions in the statistical model is false, which could lead to different substantive conclusions. This alternative does not seem to be routinely considered in meta-analysis. I believe that an important methodological challenge for meta-analysis is to develop informal or formal ways to check the assumptions such as those just listed. Checking these assumptions would give one more faith in the substantive conclusions from the formal statistical models of meta-analysis.

Statistical Inference from Models. The goal at this stage is to use the models to explain essentially all of the variability in the study results except for within-study sampling variability. That is, we want to use the model(s) developed at the previous step to model the between-study variability. For example, we might want to test the null hypothesis that the θ_i's across all studies are equal against the alternative that at least one θ_i is different. A different alternative could be that there are two different θ's corresponding to some prespecified characterization of the studies. Under the assumption that all θ_i's are equal, we might want an overall estimate and confidence interval using all the studies. If there are moderator (explanatory) variables X_i and Z_i, we might want to fit some linear or nonlinear regression model relating θ_i to X_i and Z_i. The basic statistical approach for answering such questions is to use the estimated variability $s_i(T_i)$ for each T_i to see if the hypothesis or model in question is adequate to explain all the between-study variability. Statistical procedures can be developed to do this and are discussed by Prof. Hedges.

The real problem, as in the previous step concerning model construction, is what to conclude if the null hypothesis is rejected. While the alternative being considered might in fact hold, it is also possible that one or more assumption such as those listed previously might not be valid. The challenge for meta-analysis methodology, as in other statistical applications, is to decide when the framework of a specific mathematical statistical model is useful for analysis. I would like to see more discussion and work on this issue.

Presentation of Results. It is obviously important that the conclusions be presented so that the intended audience understands and believes them. This is an important part of any statistical application. I suggest that the presentation of results from a meta-analysis include plots like those discussed earlier in the *Data Analysis* section, but supplemented with additional information from the *Models* and *Statistical Inference* stages. For example, if the between-study variability can be adequately modeled by dividing the studies into two groups with a common θ in each group, this situation can be shown on the plot by ordering the studies along the abscissa so that those in the same group are adjacent and members of the group are labeled, say using braces. Similarly, we can supplement other plots to show fitted relationships to moderator variables or confidence intervals.

The advantage of this approach is that it stays close to the data and is not likely to overwhelm the audience with confusing technical details. Technical discussion does, of course, have its place, but it is also important to present the results so that the audience finds them plausible and intuitively reasonable, given the data. Appropriately chosen plots are more likely to achieve this than are complicated tables and statements about levels of statistical significance.

In summary, in my view the problems of meta-analysis are clearly important. The key points of meta-analysis fit nicely within the standard stages of a statistical analysis. But, as with other statistical applications, doing a meta-analysis well requires much care and thought at each stage.

INTEGRATION OF EMPIRICAL RESEARCH: THE ROLE OF PROBABILISTIC ASSESSMENT

Thomas B. Feagans
Decisions in Complex Environments

1.0 Introduction

If risk assessments conducted to support important environmental decisions are to use all of the relevant information available, some means of combining or integrating empirical studies is needed. Such means should be designed and discussed with a clear understanding of the function being served by the integration. One such function is that served by a probabilistic assessment. The primary purpose of this paper is to address probabilistic assessments and the function they serve within the decision-making process.

Probabilistic assessments and their function will be clearly distinguished from two other types of integration which serve two other functions. Terms such as 'integration,' 'synthesize,' and 'combine' are ambiguous. Meta-analyses[1] And what will be called state of information assessments[2] also integrate information, but in different ways for different purposes. The functions served by these two types of integrations will be described briefly in sections 2.0 and 3.0, respectively. These descriptions serve to distinguish the other two types of integrations from probabilistic assessments, but are not comprehensive discussions of these two complex topics.

Probabilistic assessmentsand their function are described in section 4.0. Being the focus of the paper, this topic is discussed more thoroughly. The concept which underlies various possible approaches to probabilistic assessment is the concept of probability In section 4.1 probability is discussed both from the perspective of the producer and from the perspective of the user of probability assignments.

In section 4.2 an approach to probabilistic assessment is presented and advocated. The approach presented is advocated as a normative framework because it has greater generality than the alternatives.

A distinction is made in section 4.3 between advocating that an approach to probabilistic assessment be regarded as a normative framework and prescribing that an approach be applied in a particular circumstance. Two less general approaches that are special cases of the general framework are mentioned as possibilities for those circumstances where full generality is not needed and/or feasible.

There is a dual nature to the functions provided by three types of integration discussed below. On the one hand, each integrates available knowledge; but on the other hand, each deals with and represents uncertainty. Although much has been written on uncertainty, the great complexity of the topic has been underrated. The need for three types of integration of knowledge and concomitant means of dealing with and representing uncertainty has been overlooked. As a result, the unique principles and mode of thought that should guide probabilistic (risk) assessments do not seem to have been thoroughly understood. In section 5.0, the apparent reasons this situation has persisted are analyzed, and some practical reasons for improving the situation are identified.

2.0 Meta-Analysis

One means of integrating knowledge under development within the discipline of statistics, is "meta-analysis." The development of meta-analysis began in earnest when the size of the research literature on various topics which needed to be integrated for the purposes of education program evaluations, regulatory policy assessments, and other policy-related analyses became so large that narrative integrations were deemed unsatisfactory. "Although scholars continued to integrate studies narratively, it was becoming clear that chronologically arranged verbal descriptions of research failed to portray the accumlated knowledge."[3]

There are by now some standard definitions. The original analysis of data in research studies is called primary analysis.[4] Typically, statistical methods are applied in such analyses. The reanalysis of such data for the purpose of answering new questions, or the original research question with better statistical techniques, is called secondary analysis.[5]

A distinction is needed before addressing the definition of meta-analysis. Pooling the data analyzed in a set of primary analyses and deriving new results from the pooled data needs to be distinguished from a statistical analysis of the set of results of a set of primary analyses. Current definitions tend to define meta-analysis

For many situations data will be sparse for some aspects of the phenomena to be represented no matter how the model is constructed. For example, many dose-response relationships of interest extend beyond directly applicable data. The importance of simulation modeling and the generality of the probability theory are enhanced in such situations.

Model construction is a joint process requiring cooperation between those persons expert at probabilistic modeling and those persons with expertise concerning the substantive phenomena being modeled. Lack of expertise of both types can lead to avoidable mistakes. Those untrained in probabilistic modeling can make avoidable mistakes of one type and modelers unfamiliar with the substantive phenomena being represented can make another type. Both types of avoidable mistakes are possible despite the fact that there is generally no right or correct model.

4.2.2 Selection of Probability Assessors

Probability judgments are made as inputs to the probabilistic model for the most significant uncertair factors. These judgments represent both the knowledge and uncertainty about that factor. They are also produced by particular individuals with particular histories, expertises, and points of view.

Probabilistic assessment is more subjective than meta-analysis. Two probability assessors will generally not make exactly the same judgments even based on the same evidence and even if they have similar points of view. Furthermore, unless the available evidence is strong in its implications for the judgments to be made, equally well informed individuals can diverge significantly in the judgments they make.

Since probability assignments can be so subjective the choice of those individuals asked to make the judgments is very important. Thus, the subprocess of selecting who is to make the probability judgments for significant uncertainties is very important.[41] For each uncertainty this subprocess involves the steps of identifying a set of highly qualified candidates, deciding how many assessors to have and then selecting a "balanced" set. In this context balance would be a matter of representing adequately the diverse perspectives that exist within the set of highly qualified candidates. Lack of balance would be having all the judgments made by those who share a similar perspective when diverse perspectives exist.

Obviously, it is important that those with diverse perspectives be identified in the initial step of the subprocess. There will not be perfect information about perspectives and their implications for probability judgments apriori, but peers will be familiar with ways in which perspectives diverge among their colleagues, so a balanced set can be identified and chosen. The details of the participation of agencies, review committees, etc. can vary, but in one way or another the involvement of peers in the selection process is preferred.

4.2.3 Elicitation of Probability Assignments

A probabilistic assessment derives (probabilistic) implications from all of the available information and analysis relevant to the connection between possible policy alternatives and the consequences of concern. In making probability assignments probability assessors integrate diverse studies, background information, and any other considerations they deem relevant. The flexibility needed to assure that any considerations deemed relevant can be factored into the judgment goes hand in hand with the fact that there are no hard and fast rules for making probability assignments in general.

An example of the kind of judgments health experts may be asked to make is the probabilistic representation of the uncertain relationship between doses of a (suspected) pollutant and the resulting response (under specified conditions) in a specified group of people. Probability assignments may involve two uncertainties: uncertainty as to the fraction of people (if any) in the population for whom a causal relationship exists between exposure to realistic levels of the pollutant and the occurrence of a given adverse health effect; and, uncertainty as to the level that would affect a given fraction if a causal relationship indeed does exist for that fraction. In such cases the required set of probability judgments may be decomposed into two sets of judgments from which the required set may be derived mathematically: a set of probability judgments concerning existence of a causal relationship below a specified upper bound level for various fractions of the group; and a set of probability judgments concerning the level at which various fractions of the group would be affected if a causal relationship does indeed exist for the fraction addressed below the specified upper bound level. .

How to integrate various studies in making the probability assignments is a decision made by the substantive expert. Whether the probability judgments concerning a dose/response relationship are decomposed into two sets is a modeling decision to be made by the modeler, after consultation with the substantive experts. But after all such modeling decisions are made it is the substantive expert who decides how to incorporate both positive and negative studies, and how to integrate studies of various types, such as epidemiological, human clinical, and animal toxicological studies.

No matter what algorithmic analyses may have been performed, each expert making such judgments is the final arbiter for his or her own judgments and for how to arrive at them. Each expert uses and mentally integrates all information that he or she believes should have a bearing on a given judgment.[42] Each expert decides how much weight to give each piece of information and how to incorporate that information. No information is deliberately thrown away. No attempt is made to encourage close agreement with other experts.

Although those with substantive expertise concerning the phenomena being analyzed are generally the final arbiters of probability assignments, they can be aided in various ways. As with modeling, the making of probability assignments is best thought of as a cooperative process in which professional analysts elicit probability assignments from substantive experts. The assignments are elicited in such a way that they are coherent, cognitive biases are minimized, and motivational biases are discouraged.

A formal process for eliciting probability assignments has been developed by psychologists and decision analysts. This process is called probability encoding.[43] As well as applying various techniques for eliciting coherent probability assignments that either are or closely represent the assigner's judgments, the process encourages assignments that are as free of cognitive and motivational biases as is feasible. Research on such biases has been applied in developing the process.

The type of coherence achieved is dependent on the generality of the theory of probability applied. Under the deFinetti/Ramsey theory sharp probability assignments are forced; under the Koopman theory upper and lower probability assignments are allowed. Coherence under the Koopman theory is simply a matter of the entire set of probability assignments being logically consistent.

Allowing upper and lower probability assignments provides the flexibility sometimes needed. When the relevant information supporting an assignment is sparse, sharp assignments may be hard to make and unnecessarily arbitrary. The allowance of upper and lower probability assignments assures that the expert making the judgments truly discerns a difference in making his or her comparisons, no matter how weak the state of information. There may be vagueness as to where this discernible difference disappears, but this vagueness is not a problem in practice.[44] Experimental evidence evaluated by internal psychometric criteria suggests that experts perform well using this general theory.[45]

4.2.4 Computation and Presentation of Outputs

Monte Carlo simulation is required except for very simple models such as the simplest of benchmark models.[46] Statements exist in the literature on applying the Monte Carlo method to the effect that the model calculation should be iterated M times, where M varies some with the statement. The perspective being taken in such statements is that at least M iterations are needed to approximate well the distribution that would be approached in the limit were the number of iterations performed exceedingly large. But computation can require significant resources, so how many iterations should be done requires a judgment as to the marginal value of improving the approximation versus the marginal value of spending the required resources to improve the assessment in some other way.

In conventional Monte Carlo simulations each input to the model on a given iteration has either been a single assigned value or a single value arrived at by random selection from a probability distribution. Iterating the computation, with the random selections from all the input probability distributions repeated on each iteration, has given the desired probabilistic output.

In a probabilistic assessment done in full generality there are two changes from the approach just described. First, in general the input probabilities are upper and lower probability assignments rather than sharp assignments. As a result, if at least one of the inputs is in terms of upper and lower probabilities rather than sharp then the output is in terms of upper and lower probabilities rather than sharp.

Second, since in general more than one individual is making probability assignments for each significant uncertain factor in the probabilistic model, there is not a unique set of probabilistic representations of uncertainty on which to base the Monte Carlo simulation. The multiple representations of the primary uncertainties capture the phenomenon of secondary uncertainty, that there is no nonarbitrary best way of representing the primary uncertainties. Ideally, the existence of secondary uncertainty is propogated to the output of the simulation. An added dimension is used to enable an output to represent secondary uncertainty.[47] Number the N uncertainties 1,2,...,N; let r_i be the number of representations for the i^{th} uncertainty; then there are r_1 x r_2 x ... x r_N = R combinations of the probabilistic inputs to the model. Were we to iterate M times for each of the R combinations then a distribution of R points could be plotted in the added dimension for each point of the corresponding output which did not represent secondary uncertainty. When R gets large a representative sample, say S, of the R combinations can be used.

In the resulting outputs rather than a single sharp probability for an event there is a probability (risk) ribbon. The graphical representation of a ribbon has width, height, and shape rather than being a single point.[48] The width and shape of the ribbon tend to reflect the state of information supporting the probabilistic measure for that event; the ribbon will tend to be straight and thin, and thus most like a single number, when the state of information for the event in question is strong and clear; the ribbon will tend to lack integrity in this sense when the state of information is weak and amorphous. Thus, the more general output conveys useful information.

4.3 Less General Approaches

The approach to probabilistic (risk) assessment outlined in section 4.2 is the most general approach but not the only possible approach. There are two other levels of generality possible corresponding to the two other levels of generality for probability discussed in section 4.1. The three levels of generality are nested in the sense that the least general is a special case of each of the other two and the middle level (Ramsey/deFinetti) is a special case of the most general.

An important distinction concerns advocating an approach as a normative framework and prescribing its implementation in particular circumstances. There are various kinds of circumstances in which one of the less general approaches can justifiably be prescribed despite the fact that from a normative point of view this introduces bias and reduces control. For example, if those conducting the assessment and/or those using the assessment are unprepared for the greater generality the potential benefits of that generality may not occur in practice.

However, even if a less general approach is prescribed and implemented, it should be done so for the right reasons and with an understanding of why and what has been lost. Currently, less general approaches are sometimes prescribed and implemented for what appear to be partially wrong reasons. In the next section the current situation in this regard is analyzed briefly.

5.0 The Probabilistic Mode

Probabilistic assessments are based on a different set of principles and require a different mode of thought than other types of assessments. The process is not directed toward estimating a correct, true, or actual probability or set of probabilities. Rather, the process is directed toward generating probability assignments that represent the state of information about the relationship of interest as well as possible. Criteria for how well this has been done in a given case are in terms of how well the process has been conducted. How well the overall process has been conducted will be a matter of how well the subprocesses addressed in sections 4.2.1-4.2.4 have been conducted. Evaluation of particular performances of these subprocesses should be conditional on the resources it is reasonable to allocate for conducting them.

When the wrong mode of thought is adopted in thinking about probabilistic assessments unwarranted conclusions about points of practical significance can be the result. At present this type of mistake is still prevalent. The explanation seems to be that the mode of thinking appropriate for the scientific research on which probabilistic assessments should be based is carried over to the probabilistic assessments. Some aspects of thinking about scientific inference should carry over but some should not.

Perhaps the most frequent type of mistake is the procrustean forcing of probability assessments into molds apparently associated with scientific objectivity. Encouraging sharp probability judgments and expert convergence has as an objective single number probabilities for specified events. Such outputs do not reflect the states of information on which they are based. They certainly are not more objective.

The paradigm which assigns the same probability (risk) to individuals who are obviously at different risks not only suppresses secondary uncertainty but primary uncertainty as well. This paradigm has the added problem that in general important information is not integrated into the formal probabilistic assessment process.

The practical upshot of such mistakes is poorly informed decision making. Even when public decision making is poorly informed for understandable reasons, it tends to result in poor public policy.

References

1. Larry V. Hedges and Ingram Olkin (1985). Statistical Methods for Meta-Analysis. New York: Academic Press, Inc.
2. Thomas B. Feagans and William F. Biller (1981a). Risk assessment: Describing the protection provided by ambient air quality standards. The Environmental Professional 3(3/4): 235-247.
3. Gene V. Glass, Barry McGraw, and Mary Lee Smith (1981). Meta-Analysis in Social Research. Beverly Hills, CA: Sage Publications.
4. Glass, et al. (1981).
5. Glass, et al. (1981).
6. R. J. Light and P. V. Smith (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review 41: 429-471.
7. Glass, et al. (1981).
8. Hedges and Olkin (1985).
9. Glass, et al. (1981).
10. Feagans and Biller (1981a).
11. D. R. Cox and D. V. Hinkley (1974). Theoretical Statistics. London: Chapman and Hall Ltd.
12. L. L. Lauden (1973). Induction and probability in the nineteenth century, in Logic, Methodology, and Philosophy of Science, IV; edited by P. Suppes, Leon Henkin, Athanase Joja, and Gr. C. Moisil. New York: American Elsevier.
13. Augustus De Morgan. Probabilities, in Lardner's Cabinet Cyclopaedia.
14. Stanley Jevons (1877). The Principles of Science, 2nd edition. London: Macmillan.
15. L. L. Lauden (1973). Charles Sanders Pierce and the Trivialization of the Self-Corrective Thesis. Bloomington, IN: Indiana University Press.
16. John M. Keynes (1921). A Treatise on Probability. London: Macmillan. (Reprinted 1962. New York: Harper Torchbooks.)
17. Rudolp Carnap (1962). Logical Foundations of Probability, 2nd edition. The University of Chicago Press.
18. Rudolph Carnap (1955). "Statistical and Inductive Probability," The Galois Institute of Mathematics and Art. Reprinted in Baruch A. Bordy, ed. (1970). Readings in the Philosophy of Science. Englewood Cliffs, NJ: Prentice-Hall, Inc.
19. Glenn Shafer (1976). A Mathematical Theory of Evidence. Princeton University Press.
20. David Krantz. Presentation of unpublished manuscript; Thurstone Psychometric Laboratory, University of North Carolina; May 3, 1983.
21. L. Jonathan Cohen (1977). The Probable and the Provable, Oxford University Press.
22. Howard Raiffa (1968). Decision Analysis. Reading, MA: Addison-Wesley.
23. Ronald A. Howard and James E. Matheson, eds. (1983). The Principles and Applications of Decision Analysis. Menlo Park, CA: Strategic Decisions Group.
24. Feagans and Biller (1981a).
25. Rex V. Brown, Andrew S. Kahr, and Came - ron Peterson (1974). Decision Analysis for the Manager. New York: Holt, Rinehart and Winston.
26. Emile Borel (1924). Apropos of a Treatise on Probability. Revue Philosophique. Reprinted in Henry E. Kyburg, Jr. and Howard E. Smokler, eds. (1964).

Studies in Subjective Probability. New York: John Wiley & Sons, Inc.

27. Daniel Ellsburg (1961). Risk, Ambiguity, and the Savage Axioms. Quarterly Journal of Economics, vol. 75, 643-669.

28. Thomas B. Feagans (1986). Resolution of the Ellsburg Paradox. To be submitted to the Quarterly Journal of Economics.

29. Frank Ramsey (1926). Truth and probability. Reprinted in Kyburg and Smokler (1964).

30. Bruno de Finetti (1937). Foresight: Its logical laws, its subjective sources. Reprinted in Kyburg and Smokler (1964).

31. Kyburg and Smokler (1964).

32. Feagans and Biller (1981a).

33. B. O. Koopman (1940). The bases of probability. Bulletin of the American Mathematical Society, vol. 46, 763-774.

34. B. O. Koopman (1940). The axioms and algebra of intuitive probability. Annals of Mathematics, vol. 41, 269-292.

35. Paul Edwards, ed. (1967). The Encyclopedia of Philosophy, New York: Macmillan Publishing Co., Inc. and The Free Press.

36. Antony Flew (1979). A Dictionary of Philosophy, New York: St. Martin's Press.

37. T. B. Feagans and W. F. Biller (1980). Fuzzy concepts in the analysis of public health risks. Fuzzy Sets: Theory and Applications to Policy Analysis and Information Systems. Paul P. Wang and G. S. Chang (eds.). New York: Plenum Press.

38. Thomas B. Feagans and William F. Biller (1981b). A general method for assessing health risks associated with primary national ambient air quality standards. Office of Air Quality Planning and Standards, U.S. EPA, Research Triangle Park, NC.

39. Thomas B. Feagans (1986). Two types of exposure assessment. Proceedings of APCA International Specialty Conference on Environmental Risk Management. Pittsburg: Air Pollution Control Association, 1986.

40. Feagans and Biller (1981b).

41. T. S. Wallsten and R. G. Whitfield (1986). Assessing the risks to young children of three effects associated with elevated blood-lead levels. Argonne National Laboratory report AA-32 submitted to U.S. EPA Office of Air Quality Planning and Standards.

42. Bruce C. Jordan, Harvey M. Richmond, and Thomas McCurdy (1983). The use of scientific information in setting ambient air standards. Environmental Health Perspectives 52: 233-240.

43. Carl S. Spetzler and C. A. S. Stael von Halstein (1975). Probability encoding in decision analysis. Management Science, vol. 22. No. 3.

44. Thomas S. Wallsten, Barbara H. Forsyth, and David V. Budescu (1983). Stability and coherence of health experts' upper and lower subjective probabilities about dose-response functions. Organizational Behavioral and Human Performance, vol. 31: 277-302.

45. Wallsten, et al. (1983).

46. Feagans and Biller (1981a).

47. Feagans and Biller (1980).

48. Feagans and Biller (1981a).

DISCUSSION

Harvey M. Richmond, U.S. Environmental Protection Agency

My remarks will address briefly some of the history and past reviews of the ideas put forward by Thomas Feagans. I will also briefly describe the current status of risk assessment efforts sponsored by the Office of Air Quality Planning and Standards (OAQPS) involving decision analytic approaches.

As noted in Tom Feagans' paper, the approach to probabilistic (risk) assessment described in the paper was developed within the U.S. EPA's OAQPS by Tom Feagans and William F. Biller. The initial risk assessment work by Feagans and Biller occurred during the review of the ozone national ambient air quality standard in 1977. The ozone risk assessment employed the traditional (Ramsey/deFinetti) decision analytic approach to probability. During the ozone standard review, EPA requested formation of a Science Advisory Board (SAB) Subcommittee on Health Risk Assessment to review the ozone risk assessment. The SAB Subcommittee met in April 1979 and raised a number of questions about the initial decision analytic application to ozone.[1] The Subcommittee encouraged OAQPS to pursue development, but not application, of the Feagans/Biller approach. The Subcommittee also recommended that OAQPS explore alternative approaches to risk assessment that employed decision analytic and Bayesian techniques.

During the period from 1979 to 1983, OAQPS pursued development and review of several alternative approaches on risk assessment to aid decision making on national ambient air quality standards (NAAQS).[2] In parallel with the alternative approaches projects, Feagans and Biller further developed the approach used for the ozone risk assessment to the more general framework involving upper- and lower-probability assignments discussed in the paper for this conference.

In the Spring of 1981, OAQPS asked six experts in a variety of fields including statistics, decision analysis, and philosophy of science to review a report by Feagans and Biller describing their approach.[3] The reviews generally indicated that the approach advocated by Feagans and Biller was promising and merited further development, although a number of the reviewers expressed reservations about the practicality of any near-term application of the approach by EPA. One of the six reviewers, Dr. Isaac Levi, Professor of Philosophy at Columbia University stated the following,

> Feagans and Biller are exploring ways and means of avoiding the polarisation existing in current theory between Bayesians and anti-Bayesians. They suggest specifying upper and lower probabilities (i.e., intervals of probabilities) for the purpose of probability and risk assessment. This old idea, going back to Keynes and Koopman and advocated by philosophers and statisticians such as I.J. Good, H. E. Kyburg and C.A.B. Smith over 20 years ago has not been widely appreciated either by statisticians or philosophers. Yet, in my view, variants on such approaches hold the most promise for yielding approaches to risk assessment which exhibit the generality, flexibility, neutrality, and lack of arbitrariness which standard "Bayesian" and "anti-Bayesian" approaches lack.[4]

The concerns of several of the reviewers about the possible near-term application of the approach are captured by the following comments of another reviewer, Dr. David Bell, a decision analyst at Harvard University:

> It is well known that there are decreasing returns to scale with the complexity of a model to the point where you can end up worse off than no model at all. I think the models here are overly complex for the current state of applied art. I agree with the approach but I believe it is a little too much all at once. I would be happier seeing more modest goals at this point. If the report is only intended to be a look at the future or as a research document as opposed to a draft of an EPA manual then I'm content. I don't believe its realistic to expect a methodology such as this to be performed with much creedence given to it, in the next 10 years.[5]

In May 1981 the SAB Subcommittee met to review the report by Feagans and Biller and the six reviews commissioned by EPA. In a September 1981 report, the Subcommittee concluded that, "While the F/B approach may have commendable aspects as a research effort, it is not, in its present form, an implementable tool for public policy decision making...."[6] The Subcommittee also recommended that the authors publish their works in peer-reviewed journals so that others in the professional community could judge the merits of their viewpoint. The Subcommittee suggested that all material relating to upper- and lower-probabilities be considered basic research and that OAQPS should focus on standard decision analysis or Bayesian methods, using single-valued probability assignments, for developing an implementable tool.

OAQPS, following the advice of the SAB Subcommittee, has focused on development of decision analytic approaches based on single-valued probability assignments since 1981. The OAQPS risk program moved from the developmental stage to a real world application with the initiation of the lead NAAQS risk assessment project in 1983. The project, managed by Argonne National Laboratory under an interagency agreement with OAQPS, has been reviewed by EPA's Clean Air Scientific Advisory Committee (CASAC) in May 1985 and

March 1986. Probabilistic dose-response relationships were elicited from 10 nationally recognized experts (4 experts for one endpoint and 6 experts for the other) for two distinct health endpoints. Probability encoding was unnecessary for a third endpoint for which a large epidemiological data base existed and a Bayesian statistical approach was used to represent the uncertainty in the dose-response relationship. The probability encoding and dose-response aspects of the lead risk project have received generally favorable reviews from CASAC members. A final report describing the methods and results from the lead risk assessment will be released shortly.[7]

EPA is pursuing similar risk assessment efforts as part of its review of the ozone NAAQS. Efforts are underway to address both health and welfare effects associated with exposure to ozone.[8,9] Both efforts employ elicitation of expert judgment to integrate the results of different studies using standard decision analytic approaches.

While the concept of using upper- and lower - probabilities has proved to be too controversial for near-term use in the risk assessment work sponsored by OAQPS, many of the ideas, principals, and specific models developed by Feagans and Biller are being used in the current lead and ozone NAAQS risk assessment projects. It is my hope that this conference will mark another step in the constructive review, discussion, and understanding of the innovative ideas and concepts Tom Feagans and William Biller have put forth in this important area.

REFERENCES

1. Science Advisory Board Subcommittee on Health Risk Assessment (1979). Review of "A Method of Assessing the Health Risks Associated With Alternative Air Quality Standards for Ozone." Washington, D.C.: U.S. Environmental Protection Agency.

2. Thomas McCurdy and Harvey M. Richmond (1983). Description of the OAQPS Risk Program and the Ongoing Lead NAAQS Risk Assessment Project. Proceedings of the 76th Annual Meeting of the Air Pollution Control Association.

3. Thomas B. Feagans and William F. Biller (1981). A General Method for Assessing Health Risks Associated with Primary National Ambient Air Quality Standards. Research Triangle Park, NC: U.S. Environmental Protection Agency.

4. Issaac Levi (1981). Review In Six Reviews of "A General Method for Assessing the Health Risks Associated with Primary National Ambient Air Quality Standards". Research Triangle Park, NC: U.S. Environmental Protection Agency.

5. David Bell (1981). Review In Six Reviews of "A General Method for Assessing the Health Risks Associated with Primary National Ambient Air Quality Standards". Research Triangle Park, NC: U.S. Environmental Protection Agency.

6. Science Advisory Board Subcommittee on Health Risk Assessment (1981). Review of "A General Method for Assessing Health Risks Associated with Primary National Ambient Air Quality Standards". Washington, D.C.: U.S. Environmental Protection Agency.

7. Thomas S. Wallsten and Ronald G. Whitfield (1986). Assessing the Risks to Young Children of Three Effects Associated with Elevated Blood-Lead Levels. Argonne, Illinois: Argonne National Laboratory.

8. S.R. Hayes, T. Wallsten, and R. Winkler (1986). Design Document for a Study to Develop Health Risk Estimates for Alternative Ozone NAAQS. San Rafael, CA: Systems Applications, Inc.

9. Donald C. Peterson, Jr. (1986). Workplan to Develop Probabilistic Damage Functions Relating Ozone to Yield Reductions of Selected Forest Tree Species. Boulder, CO: Energy and Resource Consultants, Inc.

DISCUSSION[*]

Anthony D. Thrall, Electric Power Research Institute

First, I would like to thank Tom Feagans not only for today's interesting presentation, but also for his efforts over the past years to alert the rest of us to some fundamental problems. These problems arise when we combine the findings of various researchers in various disciplines in an attempt to make informed decisions about managing our environment. I am also glad that we have someone familiar with the workings of the EPA among our invited speakers because many issues concerning the combining of studies are likely to be specific to the field of application.

For example, the environmental field seems to differ in at least one important respect from the field of application presented by our first speaker, David Eddy. That is, the range of conditions and outcomes seems to be much more restricted in studies of a specific medical treatment than is generally the case in environmental studies. Similarly, in the study of adverse drug reactions described by David Lane, the observational condition of principal interest is the taking of a specific drug, and the outcomes of interest are whether an adverse reaction did or did not occur. In contrast, the review or establishment of an environmental policy requires us to formulate alternative policies and to consider many different types of outcome. On the other hand, the lessons learned in developing meta-analysis for making decisions about education, described by Larry Hedges, may be more directly applicable to environmental studies, since both education and the environment are complex areas of public policy.

I will not attempt to provide a technical critique of Tom's presentation in the remaining time. Instead I will simply highlight what seem to me to be Tom's main points, as an invitation to others to join in the discussion. But I would first like to try to bring us back to the present, everyday world in which we must decide environmental issues, making what use we can of all relevant information.

PURPOSES OF SYNTHESIZING ENVIRONMENTAL STUDIES

The EPA's mission is to promulgate environmental regulations that are designed primarily to protect public health, and secondarily to protect the public welfare. For the sake of discussion let's focus on public health. A suspected environmental problem goes through various stages of scrutiny by the agency. At several of these stages, multiple studies must in some way be synthesized, and therefore the methods of synthesis that have been presented deserve consideration by the agency. Here is one characterization of the various stages of scrutiny, as I have called them, with brief comments about how the methods of synthesis might be applied.

Identifying potential problems. Oftentimes one or more animal studies suggest that a compound, present at some concentration in the environment, may possibly have an adverse effect on human health at some dose. In these circumstances it may be appropriate to conduct some form of test of the statistical significance of the combined studies, as discussed by Hedges. However, rather than cautiously championing some educational innovation, we are cautiously evaluating whether we can afford to dismiss this problem for lack of evidence. (This decision, of course, would not be based on statistical significance alone.) The appropriate null hypothesis is therefore that for at least one of the sets of experimental conditions, there is a nonzero effect. In Hedges' notation, we would apply a Fisher or Tippett test, say, to the set of (1-p) values.

Estimating the magnitude of a problem, i.e., the toxicity of a compound. This can be put in the estimation framework discussed by Hedges. In education, the standardized test ("instrument") seems to define what is being measured. There may be no special attachment to the scale of the standardized test, so dividing the treatment-control difference by some standard deviation, s, to obtain an "effect size" presents no inconvenience. Moreover the division by s may be necessary if we are trying to combine the results of different standardized tests. In contrast, if we are to combine the treatment-control differences from several bioassays in a manner that is toxicologically meaningful, we may need to retain the original units or convert to a common toxicological unit. For example if the bioassays are on different species, we may need to estimate the human toxicity corresponding to the results of each study before combining studies.

Estimating public exposure to a compound. Again we have a difficult estimation problem. As Harvey Richmond has pointed out in the case of lead, we are interested in both the total exposure and the exposure that is subject to some degree of control. Total exposure might be estimated from environmental measurements, but the controllable portion of the exposure must typically be estimated by computer simulations of the emission and dispersion of the compound. Here, we might consider using expert opinions about how to combine the output from different simulation models or different model runs.

Formulating alternative regulatory actions could conceivably require us to combine multiple studies of, say, the degree of disruption that would result from various environmental controls. In practice, the formulation of alternatives seems to be less of a synthesis of multiple studies than an iterative process involving many people outside and inside the agency, and at many levels of responsibility. In any case, it should be noted that the technical or managerial staff who must synthesize various studies may be the same people who are defining or helping to define the regulatory alternatives to be considered.

Estimating the reduction in the risk to public health for each alternative. This is an extremely important and difficult estimation problem that may involve, in David Eddy's terminology, one or more chains each composed of one or more links. As discussed by Feagans, the studies to be combined at this stage are not homogeneous but rather bridge the causal span from regulatory policy to benefits in public health.

Estimating the cost, in various forms, of each regulatory alternative. Again, this could be a very complex estimation problem depending on the desired degree of realism. I do not know to what extent current practice entails multiple studies, nor am I familiar with the special problems of combining such studies. It would be illuminating to hear more on this topic from the EPA's Office of Policy Planning and Evaluation.

Making a decision. The EPA Administrator reviews and eventually approves the proposal or range of proposals that have been developed within the agency. The Administrator, then, is the reader for whom the agency's "decision package" has been developed. Undoubtedly over the course of several proposals, the reader and the authors learn how to make the decision package most useful, that is, what level of summarization or detail works best, the type of tables and graphics that are helpful, and so on. To the extent that successive administrators agree on these matters, it would seem that this is an opportunity to hone this particular process of combining studies so that the decision package fits usefully into the broader context in which the decision must be made.

Once approved by the Administrator, the proposal or range of proposals is published in the Federal Register and comments are invited from the public. It sometimes happens that these comments include studies sponsored by trade associations or environmental organizations. It can also happen that the comments prompt the agency to conduct further studies. Thus the previous integration of studies leading to the proposal must be updated. The Administrator then makes a final decision, which may be reported in the Federal Register and/or promulgated in the Code of Federal Regulations.

I wish to make two rather obvious but nevertheless important points about the role of combining studies at the decision stage. The first point is that those of us who help to prepare such "decision packages" should not forget that the package or system is merely a technical aid to the individual or group responsible for making decisions. (David Eddy's situation strikes me as unusual in this regard, for David was both the principal developer of the decision tools he describes, and a member of an advisory board only one step removed from the final decision.)

The second, related point is that there is, I believe, a natural tendency on the part of public administrators (at all levels) to ascribe responsibility for difficult highly technical decisions to some disembodied decision-making process, if it exists. I think this should be resisted. Administrators deserve credit and sympathy for having to make difficult decisions precisely because their decisions must be based on human judgment about a larger set of concerns than can be "packaged" or "processed." Moreover, as pressure is brought to bear for the process to yield a decision that at the time appears desirable or necessary (and this pressure may come more from the middle managers who are familiar with the daily evolution of technical information), the technical synthesis becomes distorted and less useful as a genuine source of information. Difficult though it is, I believe we must recognize and honor the distinct contributions of technical synthesis on the one hand and individual or group judgment on the other. Both are necessary for responsible decisions.

HIGHLIGHTS FROM FEAGANS

I would like to restate what I understand to be Tom's key points, with occasional editorial comments. Again, my purpose is to invite discussion by others.

Uncertainty and variability should not be confused. Uncertainty refers to our state of knowledge, whereas variability refers to the world around us. We are uncertain about, say, a dose-response function, i.e., the response to different doses averaged, at least conceptually, across a large population. Even if this curve of averages were known, however, it is reasonable to expect that individuals would vary about the average at each dose. (If the response is dichotomous, i.e., an individual either does or does not respond, then the average tells us the proportion of the population that responds to the given dose.) Moreover, Feagans argues that the calculus of probability, which is legitimately applied to variability in the world, should be replaced by a different calculus when it comes to uncertainty.

I agree that is important to distinguish between uncertainty and variability, but I would add that the distinction is sometimes subtle. In the example of the dose-response curve, we can imagine that technical improvements allow a more accurate determination of delivered dose and that the individual variation about the new dose-response curve is substantially reduced. The reduction in conditional variability (the variability of the responses to a given dose) can be interrupted as an increase in certainty, i.e., we are more certain of the response of an individual to the more accurately determined dose. Thus the technical advance increases our certainty by explaining some part of the variability (the part due to poorly measured dose, in the example). Of course, the basic distinction between the state of our knowledge and the variability of the world remains intact, since the overall variability of responses to the uncontrolled doses (however measured) "delivered" by the environment remains unchanged.

The goal of combining studies is to reduce uncertainty, not variability. Variability might be reduced by changing our management of the environment, but this is another matter.

There are alternative ways to quantify uncertainty. Feagans proposes that we change terms, from "uncertainty" to "degree of confirmation." I also prefer the latter term because it emphasizes the use of experimental or observational results as evidence for or against a scientific proposition, and it is less likely to be confused with some quantification of one's feeling of uncertainty. Feagans goes on to discuss three different schools of thought with regard to quantifying degrees of confirmation: classical statistics (standard errors, confidence intervals, p-values, etc.), Bayesian statistics (posterior distributions, or risk profiles to use the term of David Eddy and co-workers), and Koopman's probability intervals.

With regard to eliciting the opinion of experts on a matter subject to doubt, it seems to me that we've moved away from "degree of confirmation" back toward "uncertainty." There are two distinct questions concerning the elicitation and quantification of expert opinion: 1) is the logical foundation of the procedure sound, i.e., are the results meaningful, and 2) do decisionmakers find such results useful.

As to the technical issues of what questions to ask the experts, how to ask them, and how to encode the responses, my only comment is on the last issue. That is, the simpler encodings would seem to be more reliable, i.e., a single-valued probability seems preferable to a probability interval, which seems preferable to a prior distribution. My general concern is that these methods may give a false sense of specificity. Whatever protocol is chosen, the method can only organize and express knowledge; it cannot create knowledge. (Admittedly, this is a somewhat fuzzy distinction.) In cases where the evidence is scant or contradictory the respective procedures should yield wide confidence intervals, wide probability intervals, or "vague" posterior distributions. Even if the procedures do give such readings, we can only hope that the decisionmaker is sufficiently sophisticated to discern the simple message, "We don't know."

A coherent program of research and analysis is needed. In discussing the scientific work required to determine appropriate air quality standards, Feagans identifies three major links in the causal chain leading from environmental standard to public health benefit: the effect that adherence to the proposed standard will have on human exposure to the pollutant, the relationship between exposure and effective dose, and the consequences for public health as determined by the dose-response relationship. To ensure that research funds are put to best use, Feagans makes the very sensible suggestion that we start with the last link and work backwards. That is, we should determine what reductions in dose would be most beneficial so that the most beneficial reductions in exposure and the most appropriate environmental standards can be identified. This is tantamount to designing multiple studies, a topic that seems as important as our current topic, summarizing the evidence from multiple studies. Perhaps the design of multiple environmental studies can be discussed in a future EPA/ASA Conference.

IS QUANTITATIVE SYNTHESIS REALLY NECESSARY?

In the interests of inviting discussion I would like to close by asking whether the various quantitative methods presented really improve our understanding of environmental issues and thereby guide environmental policy. According to Feagans (or my understanding of Feagans), the research synthesis we are discussing is designed to reduce uncertainty (or at least reduce confusion by quantifying uncertainty). But of course the synthesis by itself cannot reduce variations in environmental conditions or variations among individual responses to a given set of conditions.

If the aim of research synthesis, then, is to reduce uncertainty, how much effort of this kind do the various aspects of an environmental issue deserve? It seems that on close examination any issue is fraught with uncertainties. These uncertainties may not be important, however.

Consider, for example, a different area of public policy--deciding how much to spend on the maintenance and improvement of roadways. Here it seems that reducing uncertainties is less important than reaching a compromise between adequately informed people who have different priorities. The costs of roadway maintenance must by now be well established, as are the reasons for roadway maintenance, e.g., safety, reduced wear on vehicles, promotion of commerce, and the pleasure of driving one's car on well-maintained roads. The last consideration, by the way, may be the most decisive and the most difficult to quantify. But perhaps the quantification of this factor through multiple surveys of drivers followed by a synthesis of the multiple results is unnecessary. Are there analogous factors in environmental policy?

It seems to me to be worthwhile to take stock of how we are currently piecing together the results of various environmental studies and the manner and extent to which such syntheses guide environmental policy. How, for example, did we decide to phase out leaded gasoline? Case studies of this kind would help to clarify the current and potential benefits of the quantitative methods under discussion.

1. The opinions expressed by the author do not necessarily reflect the prevailing views of the Electric Power Research Institute.

DISCUSSION

Miley W. Merkhofer, Applied Decision Analysis, Inc.

In the course of his paper, Tom Feagans has raised and commented on some very fundamental issues. Some of these issues, such as the role or function of analysis and the meaning and use of probability, are highly complex and more than a little controversial.

Since there are aspects of Tom's paper with which I disagree, it is only fair that my comments be prefaced with two confessions. First, I am a decision analyst. Like most decision analysts, when confronted with a problem requiring a subjective approach to probability I use the standard Bernoulli - Laplace - De Morgan definition of probability that Tom criticizes as the "level-two generality." It is only natural to expect that proposals requiring the relearning of basic concepts and methods of analysis would be approached by decision analysts with a sense of skepticism. Second, there has been a tremendous amount written about the various definitions of probability. As a practitioner rather than a student of the philosophy of science, I have not read most of this literature. Thus, you might characterize me as somewhat biased and largely uninformed about many of Tom's key arguments. Those of you who have followed Tom's paper will, of course, recognize this as Tom's "canonical situation"!

Having appropriately undermined my credibility in this context, I will now proceed to my comments. My discussion will be organized into three segments. First, I'll give you my view of the basic problem that Tom is addressing. Second, I'll indicate what I think are some of the important criteria for judging approaches for integrating information for environmental decision making. Third, I'll indicate some specific aspects of Tom's paper with which I agree and some with which I disagree.

As I see it, the basic problem is as illustrated in Figure 1. Health, safety, and environmental decisions involving risk must be based on scientific knowledge, but the gap between the available knowledge and the information that would make decision making easier is great. In most cases what is offered to decision makers is inconclusive data, unstructured opinion, and debate. Risk assessment (or, more generally, probabilistic assessment) is meant to provide an efficient link between scientific knowledge and decision making--a link that is designed to lead to more efficient and defensible decisions. It is in this sense that probabilistic assessment may be regarded as a means for integrating empirical research.

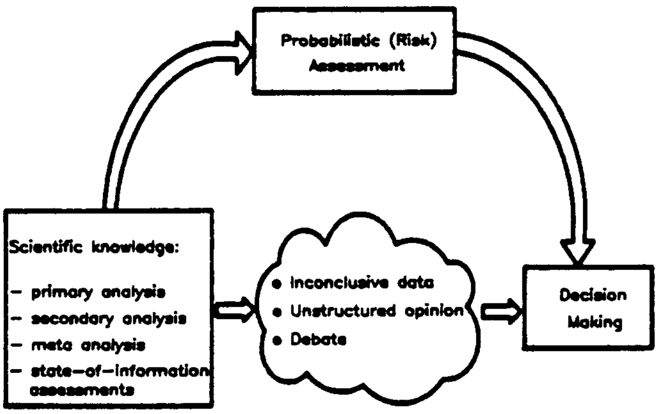


Figure 1. The Problem of Health and Environmental Decision Making and the Potential Role of Risk and Probabilistic Assessment

Figure 2 indicates the way risk and probabilistic assessment work. A model representing the cause-effect linkages between decision alternatives (such as choices among regulatory policies) and consequences (like numbers of deaths and various types of morbidity) is constructed. Because of lack of knowledge, some of the parameters of this model and the structure of the model are uncertain. This uncertainty is described and quantified using a theory of probability, and the model is then used to translate the parameter and model probabilities into probability distributions over risk outcomes. Finally, these distributions are summarized using various statistics (such as expected value) to provide quantitative indices of risk.

Figure 3 illustrates an important point about the nature of risk and probabilistic assessment. What is quantified is not real-world risk, it is risk as represented by a model, an abstraction of reality. It is clear to anyone who has ever conducted a risk assessment that the process involves making many simplifying assumptions and approximations. What is sometimes less clear is that the quantitative measures produced by a risk assessment must be translated back into the real world. The measures produced by a risk assessment are based on an imperfect, incomplete approximation. Making a decision requires considering the risk estimates along with other information relevant to the decision, including values and information about the weaknesses and limitations of the risk assessment. Thus, a risk or probabilistic assessment can never be a means for making decisions; it is only an aid to decision making.

Recognizing that risk and probabilistic assessment are aids to decision making makes it easier to identify some of the characteristics we would like the methods used in such assessments to have. Table 1 shows some evaluation criteria that would seem to be important when comparing methods. The criteria are categorized as either internal or external. Internal criteria, such as logical soundness, completeness, and accuracy, lie within the domain of analysis and relate to the quality of analysis. External criteria reflect the desires and constraints imposed by users of risk assessment, by the public, and by the limitations of time and resources.

Logical soundness relates to the degree to which a method can be justified in terms of

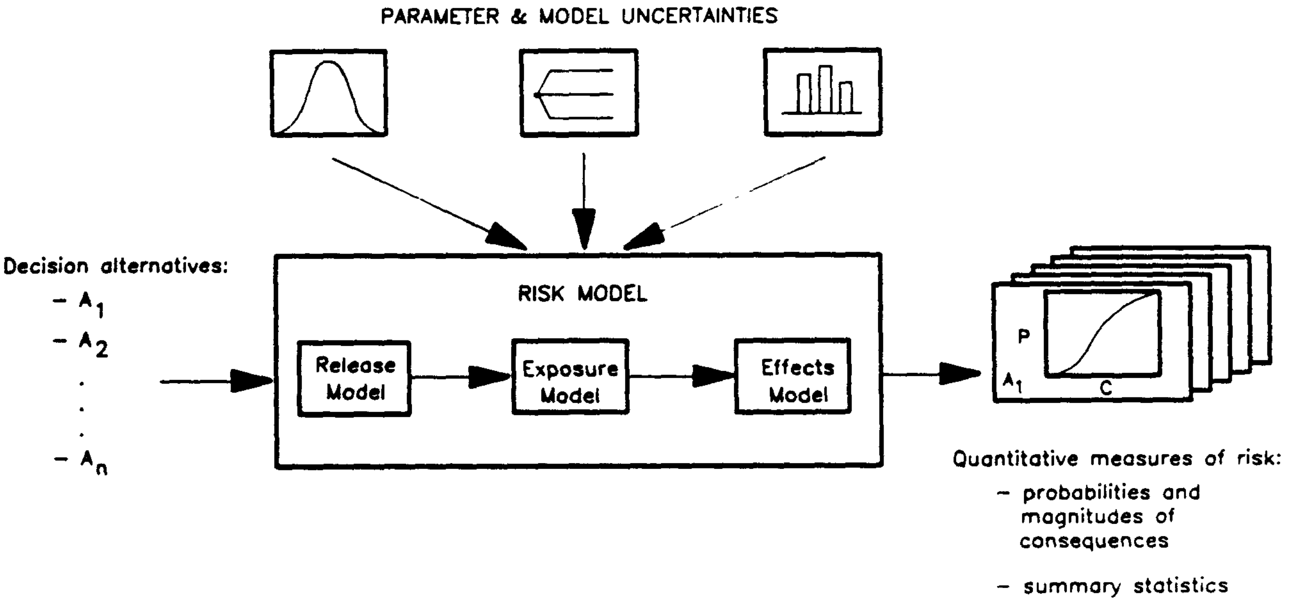


Figure 2. Generating Risk Estimates with a Risk Model

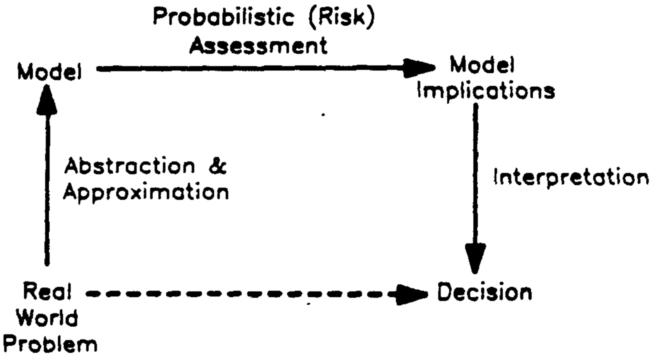


Figure 3. The Relationship Between Risk or Probabilistic Assessment and the Real World

Table 1
SOME CONSIDERATIONS FOR EVALUATING ASSESSMENTS AND ASSESSMENT METHODS

• logical soundness	internal
• completeness	
• accuracy	
• acceptability	external
• effectiveness	
• practicality	

theory and whether actual applications are likely to violate fundamental assumptions. Completeness addresses whether the method accounts for all important problem aspects and whether, due to difficulties encountered in practice, an analyst who uses the method is likely to omit certain considerations because they are difficult to accommodate. Accuracy relates to the precision and possible biases of the method and to the sensitivity of assumptions that have not or cannot be tested. Acceptability relates largely to the attitudes of and perceptions of potential users, clients, and consumers of risk and probabilistic assessment, especially decision makers and the public. Effectiveness deals with the method's ability to enable risk and probabilistic assessment to accomplish its intended ends; namely, describing and quantifying the level of risk in a way that is useful to the decision making process. Practicality reflects the extent to which the method can be conducted in the real-world, problem-solving environment using available resources and information.

The above criteria are not necessarily complete. Other sets may be preferable. The above set is offered for two purposes. One is to justify an opinion that there is no "right" or "wrong" way to define probabilities and to perform probabilistic assessments. The reason for this conclusion is that it would be unusual for any single method to be clearly superior according to every evaluation criterion. Indeed, strengths in some areas, such as logical soundness or completeness, typically lead to weaknesses in others, such as practicality. The second reason for introducing explicit evaluation criteria is to suggest that the appropriate way to judge whether one method is superior to another is not to determine whether one is preferable according to one or two dimensions, but to identify all of the dimensions that are important, estimate the performance of each method along each dimension, and then consider the relative importance of the various dimensions.

Although we have used different words, I believe that Tom shares most, if not all, of the views expressed above. Tom's paper recognizes the value of a general theory of probability as a practical means for integrating empirical studies. He places this theory within a decision analysis framework with the intent of providing an effective aid for decision making. Furthermore, he argues that "less general" definitions of probability and methods of analysis should be selected if called for by the specifics of the problem at hand. These are valid and important points that Tom contributes in his paper.

In order to explain further the logic of approach selection and to provide a basis for discussing some of the points in Tom's paper with which I disagree, it is useful to apply some of the explicit evaluation criteria introduced above. For example, the criterion of logical soundness might be applied to evaluate classical, Bayesian, and intuitive probabilities, the theories that Tom respectively refers to as level-one, -two, and -three generalities.

The classical definition of probability would seem to score high on logical soundness. The theory has a long history and is well developed. The practice of using relative frequency is supported by the law of large numbers and considerable empirical evidence. The principal weaknesses appear to be defining and justifying the conceivable outcomes to an uncertainty as equally likely and treating a small number of data points as if they were a large number of identical trials. All in all, though, classical probability is clearly a well-developed, internally consistent theory.

What about Bayesian probabilities? Tom criticizes the Bayesian approach because the standard method of elicitation does not necessarily reveal personal probabilities. His argument is that the state of information on which probabilities are based matters and that the information underlying the preference lottery used for eliciting subjective probabilities is different than that for the uncertain event. In particular, the reference lottery is a "known" probability, whereas the uncertainty is an "unknown" probability. According to Tom, when the subject says that he is indifferent between betting, for example, on a probability wheel and betting on the uncertain event, he is not necessarily equating probabilities. Like the subjects in Ellsberg's paradox, he may simply prefer to bet on known probabilities.

It is, of course, true that the state of information matters with Bayesian probabilities. This is the essense of the subjective approach. It is also true that elicitation techniques may fail to elicit probabilities that are consistent with a person's underlying information and beliefs, due to cognitive biases for example. However, this is a practical difficulty rather than a logical flaw. The reasons for this assertion follow.

For probability encoding, analysts use a variety of different elicitation techniques. They deliberately switch among frames of reference for the purpose of identifying and alerting the subject to any inconsistencies in reasoning. Betting is only one analogy that is used. The analyst will also ask whether the reference and uncertain events are judged equally probable.

Furthermore, one of the most commonly used probability encoding techniques, the interval technique, does not suffer from the problem that Tom mentions. With the interval technique, the analyst divides the range of uncertainty into regions that are judged by the subject to be equally likely. For example, if the subject thinks it equally likely that the uncertain variable will lie above or below a given value, that value is assumed to be the median of the distribution. Since all comparisons with this technique are based on the uncertain event, the subject is not required to compare known and unknown probabilities.

Since subjective probabilities are based on a well-developed, internally consistent theory, Bayesian probabilities would also seem to score high on logical soundness. The fact that the basic axioms of probability calculus apply means that analyses based on subjective probabilities are similarly well founded in theory.

What about Koopman's intuitive probabilities? Unfortunately, I have not as yet had an opportunity to explore the foundations of Tom's "level-three generality." Although Tom offers several references, Koopman's theory is not widely known, as evidenced by the fact that it is not mentioned in the dozen or so reference texts that I keep on my bookshelf. The fact that the basic axioms do not apply to Koopman's probabilities (because additivity does not hold) could be a serious problem. Theory is clearly needed to replace the important role that additivity plays in the integration and propagation of probabilities, the process that was illustrated in Figure 3.

To further explore the strengths and weaknesses of the alternative definitions of probability, the other evaluation criteria listed in Figure 4 might be usefully applied. In the interests of brevity, however, I will raise only one other point. This one relates to criteria of completeness, practicality, and effectiveness.

There is the implication in Tom's paper that, for situations where data is very sparse, subjects will find it very difficult to provide probability numbers that are sufficiently sharp for a standard analysis. I have not found this to be the case. Figure 4 provides one example. The uncertain quantity in this case is a parameter, called effective porosity, that geologists use to describe the percent of void space in rock. In this case the rock in question is a very large rock at Hanford, Washington, that is being considered as one possible location for the nation's first nuclear waste repository. The value of this parameter is important for assessing the risk that radioactive material stored in the repository will escape to the accessible environment. The curves show the probability distributions independently encoded from five experts.

Each curve in Figure 4 was carefully generated using probability encoding

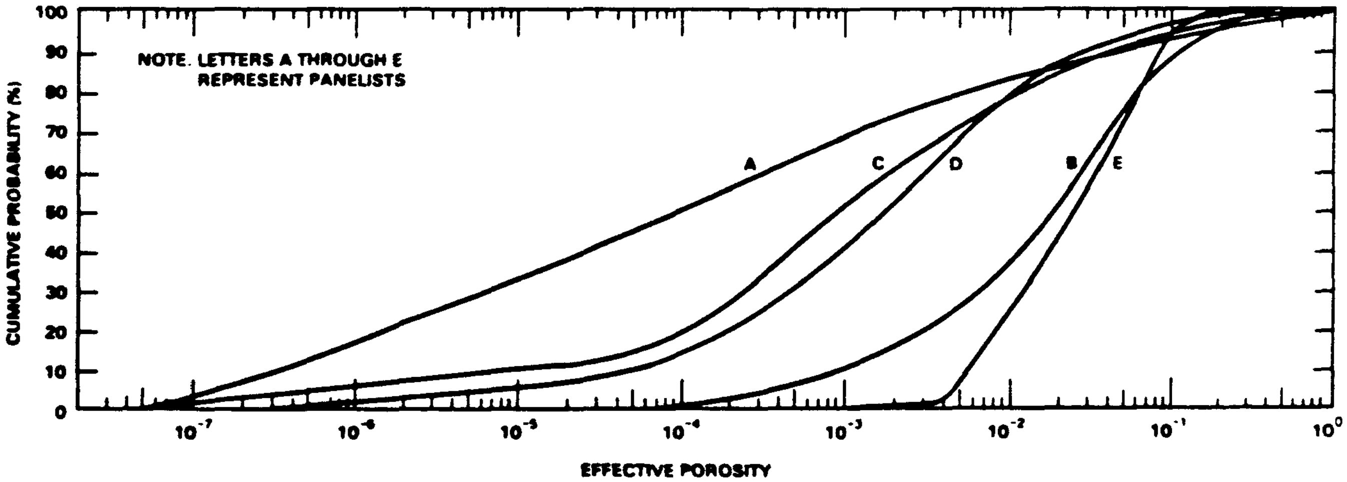


Figure 4. Cumulative Judgmental Probability Distributions for Cohassett Basalt Average Flow Top Effective Porosity at Macroscale, Obtained Independently from Experts

techniques. To support these assessments, there was, in effect, only one data point from directly applicable tests, and that test was of questionable accuracy. Therefore, it should not be surprising that the experts felt tremendous uncertainty (up to six orders of magnitude) and disagreed with one another. Despite the lack of data, the experts were quite adamant about the precise location of their curves. In every case, the final twenty to thirty minutes of each probability encoding exercise was spent exploring whether the curve should be moved two to three percentage points in one direction or the other, judgments that were in each case made definitively by the expert. The precision of these estimates was, evidently, based on convictions born of personal experience with tests on similar rock and differing theories about how processes of formation affect the parameter in question. Any uncertainty bands existing in the minds of the subjects were clearly insignificant relative to the differences of opinion. In this instance at least, the added work of assessing uncertainty bands would not have provided much additional insight to decision makers. In addition, permitting indecision in the encoding process might be less effective at forcing the sort of hard thinking that is so important to the formulation of scientific judgments.

In conclusion, although I disagree with some of the specific points of Tom's paper, I endorse his central theme that probability applied within a decision analysis framework can be a powerful and practical way of integrating empirical research. The exploration of alternative theories for the foundation of risk and probabilistic assessment is important, for advancements in this area are most likely to produce major improvements in our ability to analyze complex problems.

Risk assessment is an art as well as a science. The real challenge is to select methods that illuminate and provide insights without misleading. Research that extends the useful options or provides insights for the choice of methods is clearly of high value.

STATISTICAL ISSUES IN COMBINING ECOLOGICAL AND ENVIRONMENTAL STUDIES WITH EXAMPLES IN MARINE FISHERIES RESEARCH AND MANAGEMENT

G. P. Patil, G. J. Babu, M. T. Boswell,
K. Chatterjee, E. Linder, and C. Taillie

The Pennsylvania State University

1. INTRODUCTION

When a substantive problem needs a solution, the information needed is invariably not available as desired. Encountered or historical data may have to be used (Hennemuth, Patil, and Ross, 1986; Hennemuth, Patil, and Taillie, 1985; Patil, 1984; Patil, Rao, and Zelen, 1986). Often, an ad hoc decision is made by the manager based on incomplete or inadequate data involving similar situations, perhaps augmented by various experts opinions. Ecological studies in this manner have been done on a continual but informal basis. This points up the importance of developing systematic methods to combine studies. Three approaches that have been generally used are: (1) combining different data sets to obtain a long-enough time series or a large-enough data set to perform the desired analysis; (2) combining the results of different studies; and (3) combining expert opinions.

1.1 Combining Data Sets

Usually, pooling occurs for the same type of data taken under different conditions, including different locations, and different seasons or different years. Alternatively, entirely different types of data may be combined. It becomes necessary to assume that some underlying common features exist among various data sets; in order to extract these features, it is necessary to transform the data sets to make them comparable.

Section 2 is an example where individually small recruitment data sets, (giving the number of fish of "catchable" size entering a fish stock), for different species of fish and different stocks from various oceans are combined to give a data set large enough to estimate a "universal" recruitment distribution. This may then be used to estimate a recruitment distribution for an individual fish stock.

Another common situation arises when a change in the instrumentation or in the data collection protocol occurs during the course of an investigation. If there is only one such change, then the two data sets need to be combined into one. Here the purpose of combining data is to obtain a consistent data set for use in testing hypotheses, investigating trends, etc. Section 3 is an example that involves a change from one ship to another in a marine fisheries research trawl survey. A paired experiment is carried out to compare the fishing power of the two ships. The results of the experiment are used to calibrate the two data sets.

1.2 Combining Results

Extrapolation from one situation to another is usually done by assuming a super model that combines the results of the two different situations.

Section 7 describes a method for assessing the risk of a toxicant to a species of fish by utilizing results from laboratory tests. Estimates of toxic concentrations obtained from various bio-assay tests are combined to form new data sets. A pattern is established for each of the data sets by curve fitting. The curves are used in turn for extrapolating the long-term toxic effect as a function of the short-term effect on a species for which the results on short-term effects are available.

Sections 4, 5 and 6 discuss an environmental index effort for coastal and estuarine degradation. The overall approach combines the results of a statistical analysis on data from a control region with the data from a test region to produce a single number indicating in some manner the health of the environment. Section 5 uses the reproductive success of osprey as an example. The reproductive success varied in time with DDT pollution. A period of time with little pollution effect is used as a control instead of a control region.

Section 6 is another example in which dissolved oxygen measurements are used. The formulation is different since the measurements are combined with the results of three statistical studies on laboratory experiments. The laboratory studies produce three different dose-response curves for three different species and for three different responses. The dose is exposure to low dissolved oxygen episodes. The three responses are mortality, reduced growth and avoidance. The results of these studies are combined with dissolved oxygen measurements to produce a single number for each low dissolved oxygen episode.

Section 8 describes the initial stages of an analysis whose goal is to partition the causes of early-life-mortality of fish among various climatic and pollutant variables. The data set (Summers et al. 1984) is a historical data set which combines several ecological studies. The fish data set is a stock index data set for each species. This index is the catch per unit effort which was estimated from many sources. A model relating fishing in the Potomac river and fishing in the Chesapeake Bay was constructed and the results were combined with landings data to give the stock index. The environmental data consisted of river flow and temperature. Pollution data include gross indicators of pollution such as population size, employment levels, sewage discharged, dredging and some loading variables such as dissolved oxygen and nutrients. All of these data sets were combined to give a multivariate time-series data set for analysis and interpretation.

The purpose of the analysis and interpretation of this combined data set was to be able to evaluate different statistical techniques. Furthermore, different techniques use different modeling assumptions and, therefore, may reveal different pollution effects on different species. It should be worthwhile to be able to combine the results of these analyses by the methods of combining expert opinions to provide a more accurate picture of pollution effects.

1.3 Combining Expert Opinions

Included in this approach is the combining of different models where specific p-values are available for each model. These correspond to probabilities of various propositions as related by experts. There may be two or more propositions of which exactly one is correct. Each model or expert gives probabilities of these propositions. There are two cases. Either the probabilities add to one, or the probabilities add to a value less than one. The latter case allows the possibility of an opinion to be held in 'reserve.'

Section 9 gives a brief review of some of the problems and possible approaches for the combination of expert opinions. Sections 4, 5, and 6 are also relevant in this connection. To begin with, opinions are solicited on potentially informative variables. Even after this is accomplished, there remains a considerable need to identify and utilize expert opinions on issues, such as, what data sets are suitable, how to organize the data, what to use as control and as test regions, and how to combine and summarize the data into indices that are comparable. "The choice of variables is not easy and usually involves extensive exploratory data analysis" (O'Connor and Dewling 1986).

The candidate data sets result from many different studies on different aspects of the ecosystems. These are expected to yield information regarding the health of the ecosystem. Examples include benthic species composition and abundance, fish and shellfish diseases, fecundity in fish and shellfish, mortality in eggs and larvae of fish and shellfish in the field and reproductive success in marine birds. Also included are measures of pollution, such as toxicants in marine foods, pollutants in the sediments and dissolved oxygen in the water column. When the available indices are considered over a region and through time, a picture of the ecological health of the region begins to emerge. At this stage a meta analysis of the separate index values over time and/or space should be of some help to the managers in their task of managing natural resources.

2. RECRUITMENT DATA AND KERNEL APPROACH

2.1 Background

For several years the Northeast Fisheries Center (NEFC) has been assembling recruitment series for a large number of oceanic fish stocks. Recruitment is defined by the number of fish of 'catchable' size entering a fish stock. Estimation of recruitment distributions is important for the assessment and prediction of long term frequencies of good and poor year classes. In this connection, several parametric distributional models have been fitted to each of the available recruitment data sets (Hennemuth, Palmer, and Brown, 1980; Patil and Taillie, 1981). The small sample sizes prevented reliable assessment of goodness-of-fit. It also proved difficult to effectively discriminate between competing models, e.g., between the gamma and the lognormal distribution.

In view of the preceding, Richard Hennemuth of NEFC suggested tnat the recruitment data for the various stocks be combined into a single large data set and analyzed with the two-fold purpose:

(i) better assess the fitting performance of the different methods and models,

(ii) arrive at a fairly precise estimate for a "universal" recruitment distribution.

2.2 Combining the Data Sets

Recruitment series for 18 stocks were selected for analysis. The data and histograms for the individual stocks appear in Table 2.1. Sample sizes range from 10 for North Sea mackerel to 43 for Georges Bank haddock. On the whole, the data exhibit strong positive skewness with the occasional occurrence of large positive values corresponding to the appearance of a strong year class.

When combining data, the various data sets must have some common features (or there would be no reason to combine) as well as some differences (or the matter would be trivial). The trick is to model the common features and to suitably adjust the data for the differences before combining. The large combined data set is then used to draw reliable inferences concerning the common features.

In our case, it is hypothesized that the p*th* recruitment data set can be described as a random sample from a scale-parameter family of distributions

$$F(x,\theta_p) = F(x/\theta_p). \qquad (2.1)$$

Here the scale parameter θ_p is allowed to vary from stock to stock. The functional form of the cdf F is assumed to be the same for all stocks and therefore represents a "universal" recruitment distribution. The p*th* data set is adjusted by dividing through by a suitable scale statistic. The arithmetic mean (divided by 5) was used in the present analysis but it may be worthwhile to mention some other possibilities:

a) As pointed out previously, large positive values are sometimes encountered. For a given stock, it may be entirely a matter of chance whether such a value occurs in the available data. The arithmetic mean is sensitive to the presence of large values. Thus, using the arithmetic mean to descale, introduces considerable extraneous variability into the combined data set. A scale statistic such as the geometric mean may be preferred for this reason.

b) The assumption that the different stocks differ only in the scaling is only an approximation. One might attempt to develop data transformations that would adjust for differences in distributional shape as well. For example, the z-score of the logged data adjusts for scaling and also for certain types of shape parameters (e.g., Weibull, lognormal).

2.3 Estimating the Universal Recruitment Distribution

Having combined the descaled recruitment values, the next step is estimation of the common cdf F. Here a nonparametric approach has been adopted. In passing, it may be noted that the problem would be trivial if we were prepared to assume a parametric form for F. In fact, if $F(\cdot) = G(\cdot,\underline{\varphi})$ where G is a <u>known</u> distribution and φ is a vector of unknown parameters, then from (2.1) the p*th* data set is a random sample from $G(x/\theta_p,\underline{\varphi})$. From this, the joint likelihood can be written down and parameters estimated in the usual way.

The nonparametric method employed is a variation of the kernel technique. The distribution to be estimated is approximated by a mixture of lognormal distributions. There is one lognormal (known as a <u>kernel</u>) for each available observation. Each kernel is "centered" so that its geometric mean is located at the corresponding observation. The various kernels are taken to have the same logarithmic standard deviation, which is known as the <u>bandwidth</u>.

The central theme in application of kernel methodology is determination of suitable bandwidth. Overly small bandwidths yield estimated pdf's whose graphs have a rough, jagged appearance. Excessively large bandwiths smooth the probability mass over a wide interval, losing most of the local features of the data. There is extensive literature on kernel methods (Wertz and Schneider, 1979) and we will not here dwell upon the technical aspects except to point out that bandwidth determination was done through cross-validation.

The histogram of the combined data set is shown in Figure 2.1. Superimposed are the fitted lognormal distribution, the fitted gamma distribution and the kernel estimate. Parameter estimation for the gamma and lognormal was done by the method of maximum likelihood. Inadequacy of the lognormal fit is readily apparent. The kernel fit, while generally acceptable, does exhibit a leftward bias for small year classes. We have been able to remove most of this bias by either of two techniques: (i) variable bandwidths and (ii) regression toward the mean of the kernel centers. The gamma distribution also shows a leftward bias, and has a right hand tail that is much too short to give an adequate fit.

2.4 Estimating Individual Recruitment Distributions Interpretation of the Kernel Estimator

The simplest estimate of the recruitment distribution for a particular stock is formed by rescaling the universal recruitment distribution. However, the assumptions which led to the universal curve are only approximations and the preceding estimate will be inaccurate in certain respects. This raises the question of whether the limited data that is available on a particular stock can be used to improve the estimate. Here the James-Stein (1961) paradigm may offer some guidance. Envision the separate (descaled) recruitment distributions as forming a cloud of points in the space of all probability distributions. The universal curve estimates the center of this cloud. Use the available data to obtain, perhaps by the kernel method, a low quality estimate $\hat{F}_p$ for a particular distribution. For the final estimate, use a convex linear combination of the imprecise estimate $\hat{F}_p$ and the precise but inaccurate universal estimate.

It may be of interest to close this section with an interpretation of the kernel estimator. The recruitment process is governed by many factors, both environmental and biological. Currently there is little understanding of what these factors are, how they operate quantitatively and how they interact. The kernel method attempts to account for the annual variability in recruitment without developing a detailed explanatory model. Consider the multidimensional space of all relevant factors and let this space be partitioned into N subsets, one for each available recruitment value; the subsets occur with the same long term relative frequency of 1/N. Conditional upon a particular partitioning set, there is still residual environmental variability within that set and a corresponding variability in recruitment. It is this variability that is represented by the lognormal kernels. Each kernel is centered at the corresponding observation, in effect treating each observation as typical for its partition set.

Also note that the bandwidth expresses the within-partition-set variability. In particular, the bandwidth must decrease toward zero as the partition becomes finer (i.e. as the number of observations increases). It follows that the bandwidth cannot be treated as a universal constant: the bandwidth appropriate to a particular stock is larger than the bandwidth obtained for the combined data.

TABLE 2.1: RECRUITMENT DATA

```
GBCD  GEORGES BANK COD
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
16        5.00       24.31        6.21    .256   -.042   2.201   3.156    .271
         37              7.400
         31              6.200
         26              5.200
         26              5.200          █
         28              5.600          █
         28              5.600          █
         25              5.000        █ █
         29              5.800        █ █
         23              4.600        █ █
         16              3.200        █ █
         15              3.000        █████
         16              3.200    +----+----+
         28              5.600    0    5   10
         18              3.600
         16              3.200
         27              5.400

NSCD  NORTH SEA COD
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
15       50.00      232.95      141.85    .609    .984   3.178   5.271    .602
        104              2.080
        234              4.680
        222              4.440
        315              6.300      ██ █
        283              5.660      ██ █
         92              1.840      ███████ █ █
         87              1.740    +----+----+----+
        368              7.360    0    5   10   15
        450              9.000
         83.2            1.664
        160              3.200
        145              2.900
        245              4.900
        124              2.480
        582             11.640
```

```
NACD  NORTHEAST ARCTIC COD
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
16      200.00     1011.31      749.99    .742   1.044   3.025   6.638    .774
        507              2.535
       1163              5.815      █  █
       2364             11.820     ███ █
       1931              9.655    ██████ █ █ █ █
        264              1.320   +----+----+----+
        172               .860   0    5   10   15
        300              1.500
        607              3.035
       1540              7.700        1024              5.120
       2782             13.910         419              2.095
        820              4.100         725              3.625
       1031              5.155         532              2.660


GBHD  GEORGES BANK HADDOCK █ █ █ 1974 - 1976 EXCLUDED
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
43       15.00       73.36       72.67    .991   3.540  19.651   3.702   1.539
         42              2.800
         45              3.000
         56              3.733
         61              4.067        █
         60              4.000        █
         57              3.800    █   █
        107              7.133    █  ██
         77              5.133    █  ██
         64              4.267    █ ███  █
        112              7.467    █ ███ ███
        116              7.733    ██████████
         63              4.200    ██████████                    █
         24              1.600   +----+----+ ---+----+----+----+----+
         65              4.333   0    5   10   15   20   25   30   35
         92              6.133
         93              6.200
         60              4.000
         34              2.267
        129              8.600
         58              3.867
        110              7.333
         49              3.267
        146              9.733
         64              4.267
        100              6.667
         78              5.200
         73              4.867
         61              4.067
        133              8.867
        127              8.467
         57              3.800
         41              2.733
        148              9.867
        464             30.933
         36              2.400
          9               .600
         11               .733
          .3              .020
          1               .067
          5               .333
          .173            .012
         10               .667
         16              1.067
```

```
NSHD  NORTH SEA HADDOCK
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
18      200.00     1082.94     1499.05   1.384   2.401   8.391   6.219   1.274
        142               .710
        632              3.160
       3005             15.025
         68               .340    █
         63               .315    █
        147               .735    ██ █
        767              3.835    ██ █  █
       6296             31.480    █████ █     █  █               █
        386              1.930   +----+----+----+----+----+----+----+
        111               .555   0    5   10   15   20   25   30   35
        901              4.505
       1324              6.620
        256              1.280
       1278              6.390
       2557             12.785
        302              1.510
        577              2.885
        681              3.405

NAHD  NORTHEAST ARCTIC HADDOCK
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
17       50.00      273.88      345.56   1.262   2.690  10.125   5.037   1.079
        479              9.580
        150              3.000
        364              7.280
        438              8.760     █
         30               .600    ██
         26               .520    ██ █    █
        247              4.940    █████ ████                    █
        140              2.800   +----+----+----+----+----+----+----+
       1523             30.460   0    5   10   15   20   25   30   35
        408              8.160
         89              1.780
         73              1.460
         89              1.780
        185              3.700
        317              6.340
         55              1.100
         43               .860

GBHR  GEORGES BANK HERRING  █ █ █  1975 - 1976 EXCLUDED
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
12      350.00     1743.25      995.86    .571    .960   2.848   7.309    .553
       3544             10.126
       1868              5.337
       2113              6.037
       1585              4.529      █
       1755              5.014      █  █
       1910              5.457      █  █    █
       1184              3.383      █████   █
        855              2.443   +----+----+----+
        759              2.169   0    5   10   15
       3844             10.983
        757              2.163
        745              2.129
```

```
NSHR  NORTH SEA HERRING
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
18     2000.00     7771.11     4620.03    .595   1.566   5.264   8.800    .571
      21370             10.685
       5640              2.820
       7820              3.910
       1980               .990       ██
      16720              8.360       ██
       7330              3.665       ██
       8730              4.365      ████
      10950              5.475     ██████  █ █
       5710              2.855    +----+----+----+
       5290              2.645    0    5   10   15
       7580              3.790
       7620              3.810
       3820              1.910
       9060              4.530
       7110              3.555
       5010              2.505
       2240              1.120
       5900              2.950

NWHR  NORWEGIAN SPRING SPAWNING HERRING
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
20     3000.00    13734.95    18576.43   1.352   2.307   7.919   8.730   1.341
      78267             26.089
      20718              6.906
      11254              3.751
      12642              4.214     █
       4680              1.560     █
       3114              1.038     ██ █
       4558              1.519     ██ █
       3723              1.241     ██ █
       2937               .979     ██ ████  █     █          █
      47442             15.814    +----+----+----+----+----+----+
      28631              9.544    0    5   10   15   20   25   30
       9927              3.309
       2807               .936       11194              3.731
      17957              5.986         687               .229
      11426              3.809         599               .200
        942               .314        1194               .398

GBMC  GEORGES BANK MACKEREL
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
16      400.00     1934.88     1776.28    .918   2.137   7.550   7.243    .792
        917              2.293
        428              1.070
        429              1.073
        541              1.353      █
       1208              3.020      █  █
       3179              7.948      ████  █
       7791             19.478      █████ ██           █
       3085              7.713    +----+----+----+----+
       3208              8.020    0    5   10   15   20
       1616              4.040
       1686              4.215
       1202              3.005
       1868              4.670
       2300              5.750
        800              2.000
        700              1.750
```

```
NSMC  NORTH SEA MACKEREL  █ █ █  1979 EXCLUDED
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
10      100.00      743.20      956.02   1.286   2.253   6.790   6.017   1.080
       1077             10.770
       3481             34.810
        635              6.350
        467              4.670     █    █
        173              1.730     ██ ████   █                       █
        524              5.240    +----+----+----+----+----+----+----+
        587              5.870    0    5   10   15   20   25   30   35
        318              3.180
         85               .850
         85               .850
```

```
NSSA  NORTH SEA SAITHE
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
18    50000.00   265372.11   153240.17    .577   1.325   4.622  12.330    .580
      60818              1.216
      80890              1.618
     196266              3.925
     141893              2.838        █
     191599              3.832        █
     154993              3.100        ███
     424108              8.482      █ ███  █
     436820              8.736      █████  ██    █
     469071              9.381    +----+----+----+
     237653              4.753    0    5   10   15
     236391              4.728
     240269              4.805
     281607              5.632
     710445           . 14.209
     255169              5.103
     179341              3.587
     196909              3.938
     282456              5.649
```

```
NSWH  NORTH SEA WHITING
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
16   300000.00  1333274.00   630841.20    .473    .514   2.271  13.981    .515
    1495430              4.985
     355120              1.184
     680024              2.267
     774709              2.582       █
     975353              3.251       █ █
    2609047              8.697       █ █
     859892              2.866       █ ██ █
     776350              2.588      █████ ██
     824927              2.750    +----+----+
    1784215              5.947    0    5   10
    2321951              7.740
    1606143              5.354
    2240953              7.470
    1332680              4.442
    1441585              4.805
    1254005              4.180
```

```
SAPD  SOUTH AFRICAN PILCHARD █ █ █  1976 EXCLUDED
(I2,3F12.2,5F8.3)
 N          1X         MEAN          SD      CV    SKEW    KURT LOGMEAN    LOGSD
26     3000.00     13673.46    10153.52    .743   1.364   4.043   9.272     .714
      12500              4.167
       9320              3.107
       6750              2.250
       6060              2.020       █
       9730              3.243       █
      21300              7.100       █
      37600             12.533     ███
      41600             13.867     ███
      31300             10.433     ████ █
      26000              8.667    █████████ █ ██
      18500              6.167    +----+----+----+
      16200              5.400    0    5    10   15
      11600              3.867
       6870              2.290
       3540              1.180
       2280               .760
       4620              1.540
      11500              3.833
       9220              3.073
       6590              2.197
       5820              1.940
       9100              3.033
       4310              1.437
      11200              3.733
      18100              6.033
      13900              4.633
```

```
SAAN  SOUTH AFRICAN ANCHOVY
(I2,3F12.2,5F8.3)
 N          1X         MEAN          SD      CV    SKEW    KURT LOGMEAN    LOGSD
13       11.00        56.03       14.81    .264    .685   3.088   3.992     .260
         39.1            3.555
         33              3.000
         56              5.091
         68              6.182          █
         50              4.545        ███
         56              5.091        ████
         61              5.545        ████ █
         46              4.182    +----+----+
         43.2            3.927    0    5    10
         90.1            8.191
         54              4.909
         76              6.909
         56              5.091
```

```
SARH  SOUTH AFRICAN ROUND HERRING
(I2,3F12.2,5F8.3)
 N          1X         MEAN          SD      CV    SKEW    KURT LOGMEAN    LOGSD
13         .60         2.83        2.15    .758    .925   2.481    .751     .771
          1.3            2.167
          1.5            2.500
          2              3.333
          3.2            5.333       █
          7.4           12.333       █
          6.3           10.500      ███
          5.3            8.833      ███ ██ █ █ █
          1.3            2.167    +----+----+----+
          1.4            2.333    0    5    10   15
           .7            1.167
           .6            1.000
          3.9            6.500
          1.9            3.167
```

```
GBSH  GEORGES BANK SILVER HAKE
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
22      250.00     1207.05      810.48    .671   1.116   3.361   6.885    .647
           339            1.356
           412            1.648
           571            2.284
           883            3.532        █
          1305            5.220        █
          1993            7.972        █
          2207            8.828       ██ █
          2993           11.972       █████ █
          3257           13.028       ████████  █ █
          1951            7.804     +----+----+----+
          1119            4.476     0    5   10   15
           626            2.504
           603            2.412
           559            2.236
           517            2.068
           439            1.756
           894            3.576
          1174            4.696
          1100            4.400
          1663            6.652
          1400            5.600
           550            2.200
```

```
PVAN  PERUVIAN ANCHOVY
(I2,3F12.2,5F8.3)
 N          1X        MEAN          SD      CV    SKEW    KURT LOGMEAN   LOGSD
16       60.00      307.56      141.00    .458    .061   2.040   5.588    .589
           332            5.533
           237            3.950
           183            3.050
           403            6.717         █  █
           193            3.217         █  █
           439            7.317        ██ ██
           383            6.383      █ ██ █████
           338            5.633     +----+----+
           377            6.283     0    5   10
           553            9.217
           539            8.983
            52             .867
           160            2.667
           180            3.000
           160            2.667
           392            6.533
```

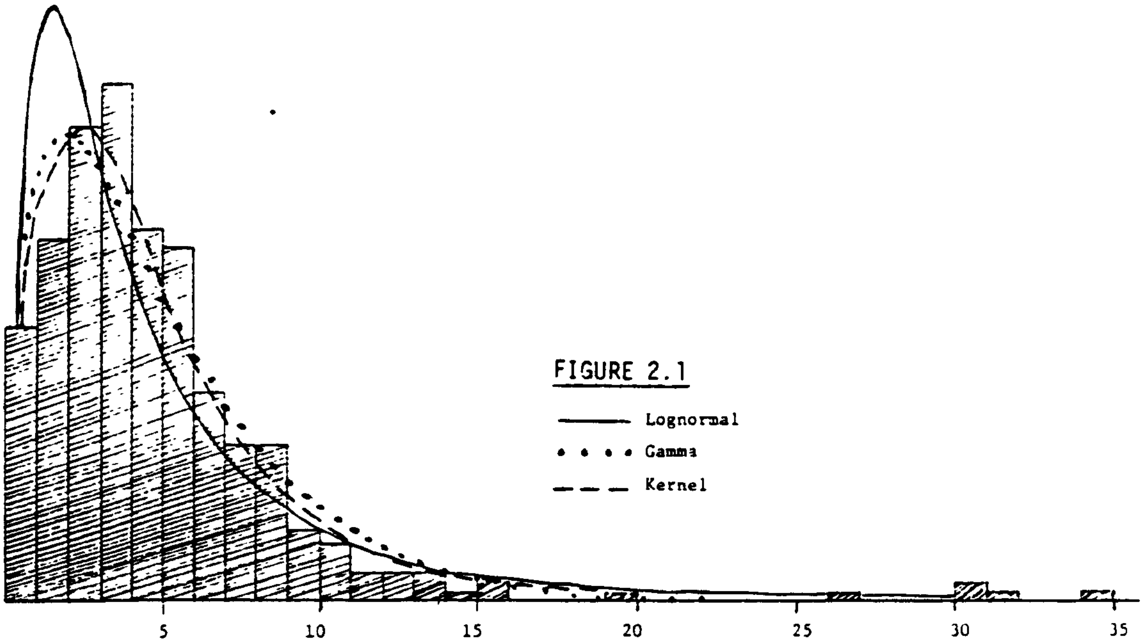


FIGURE 2.1

3. ESTIMATION OF RELATIVE FISHING POWER OF DIFFERENT VESSELS

3.1 Introduction

This section discusses standardization and pooling of data from different parts of a sample survey. Trawl surveys carried out by NEFC at Woods Hole are used to monitor the year to year changes in the abundance of several marine fish stocks. The principal objective of these surveys is to provide data necessary to assess the production potential of traditional and underutilized species (see Byrne and Fogarty 1985). These surveys may also be of use in assessing the long-term effects of pollution, where a time series of data is necessary to determine trends.

A critical aspect of any long-term survey program is the standardization of survey units. Inherent differences in vessels, nets, etc., which change from time to time, may introduce bias due to differences in the resulting fishing power. Two ships, the Albatross IV and the Delaware II, have been used at different times in the last two decades. A conversion factor may be necessary to make the various parts of the survey comparable. The conversion factor may be different for different species depending on their size, weight, schooling behavior, etc.

To see if there is any difference in the fishing power, and to estimate the conversion factor, if necessary, paired tows were made using the Albatross IV and the Delaware II off southern New England and on Georges Bank. The station locations were preselected using a stratified sampling scheme. A total of 142 successful pairs of tows were performed with these vessels during 1982 over a large area that encompassed a variety of depth and bottom types.

Byrne and Fogarty (1985) carried out an analysis of the data using non-parametric methods by rank transforming the observations. But, this method loses much of the information contained in the data.

Due to the highly skewed nature of the distribution of the catches, the difference in the mean catches is not an efficient estimate of the relative fishing power. In multispecies fish surveys, when large areas are sampled, any particular species usually occupies only a part of the total survey area. In these circumstances, the zero values can be taken to represent areas of unsuitable or unoccupied habitat. The proportion of non-zero values in the sample estimates the proportion of the total survey area that is occupied by the species.

The interpretation of the proportions of non-zeros in a sample as an estimate of habitat area may be vague in some situations, especially for mobile populations. A suitable habitat may change from time to time due to many factors including the timing of the survey, or the non-occupancy of an area simply because of low population level. However, keeping the zeros separate often enables one to fit a relatively simple distribution like the lognormal to the non-zero values.

Transformations like log(a+x) have been suggested to avoid the problem of zeros in the log transformation, where $a > 0$ is a constant. The problem here is the choice of a. The transformation using $a = 1$ has been studied at Woods Hole to transform the data to normality. Because of the large proportion of zeros, log(1+X) is far from normally distributed, which makes it difficult to retransform and interpret the results expressed in the transformed scale. Further it has been observed that different values of a near 1 lead to different conclusions so this class of transformations leads to unreliable conclusions.

3.2 Method

It is reasonable to assume that the population mean of the catch per tow varies in proportion to the relative abundance of fish over a region. Further, a zero catch by both the vessels at a station is non-informative with regard to the relative "fishing power" of the ships. Zero catches may simply be due to lack of fish in the area. Consequently, it is enough to consider those pairs of the data where at least one component is non-zero. This leads to the consideration of the independent vectors $(X_1,Y_1),\dots,(X_p,Y_p)$ where for each i, $X_i \geq 0$, $Y_i \geq 0$ and $X_i + Y_i > 0$. It may be reasonable to define $\theta = E(X|X>0)/E(Y|Y>0)$ as the relative fishing power (or the conversion factor). A natural estimate of θ is

$$\hat{\theta} = \left[\frac{1}{n_X}\sum_{i=1}^{p} X_i\right] / \left[\frac{1}{n_Y}\sum_{i=1}^{p} Y_i\right]$$

where n_X and n_Y are the number of non-zero observations X_i and Y_i respectively. If the X_i and Y_i are assumed to have lognormal distributions, then some modification of this formulation is required (see Babu 1986). Further, the bias can easily be estimated using the paired data and shown to be practically negligible. The paired data set is also used in estimating the standard errors.

3.3 Results

A total of 32 species were identified for the analysis. The non-zero values are approximately lognormally distributed. Overall relative fishing power was computed for catch in numbers and in weight. Table 3.1 gives the estimates.

Table 3.1

Catch in	Log Fishing Power	Standard Error
Weight	-.2780	0.0850
Number	-.1401	0.1040

Both in terms of total numbers and total weight, Delaware II appears to have significantly more fishing power than Albatross IV.

The results for the 32 species are presented in Tables 3.2 and 3.3. For additional discussion, see Babu, Pennington, and Patil (1986).

TABLE 3.2

CATCH IN NUMBERS

NO.	SPECIES	BOTH NONZERO			A.NONZERO D.ZERO			A.ZERO D.NONZERO						
		#PAIRS	MEAN	STAND ERROR	#PAIRS	MEAN	STAND ERROR	#PAIRS	MEAN	STAND ERROR	DIFF. MEAN	STAND. ERROR	ESTIMATE	STAND. ERROR
1	Smooth Dogfish	10	.1869	.2503	6	.3466	.1987	8	.2878	.2187	.0588	.2955	.1334	.1910
2	Spiny Dogfish	33	-.2914	.2437	10	1.2035	.8345	11	.5395	.3167	.6640	.8925	-.2251	.2351
3	Winter Skate	43	-.0954	.2178	8	1.0165	.4751	11	1.4617	.5940	-.4452	.7606	-.1219	.2094
4	Little Skate	55	-.1732	.1051	21	.6284	.2453	17	.5463	.1925	.0821	.3118	-.1472	.0996
5	Silver Hake	95	.2846	.1011	16	1.4740	.4985	13	1.1121	.4659	.3619	.6823	.2863	.1000
6	Atlantic Cod	12	.0376	.2388	5	.8318	.5133	3	1.1755	.9456	-.3437	1.0759	.0197	.2331
7	Haddock	10	-.1738	.2817	3	.0000	.0000	6	.8540	.4666	-.8540	.4666	-.3555	.2411
8	White Hake	8	.5918	.3390	12	.6550	.2583	12	1.2006	.4923	-.5456	.5560	.2835	.2895
9	Red Hake	36	.1978	.1921	13	.5713	.2705	19	.6264	.2003	-.0551	.3366	.1357	.1668
10	Spotted Hake	16	-.3107	.1531	8	.3385	.2431	8	.7758	.4931	-.4373	.5498	-.3198	.1475
11	American Plaice	5	-1.0456	.6390	5	.8254	.6967	1	2.0794	.0000	-1.2540	.6967	-1.1408	.4709
12	Summer Flounder	29	.0945	.1339	11	.7750	.2968	5	.1386	.1381	.6364	.3273	.1722	.1239
13	Fourspot Flounder	53	-.0468	.1116	11	.3151	.1665	16	.7160	.3569	-.4009	.3939	-.0731	.1074
14	Yellowtail Flounder	43	-.3764	.1065	5	.7167	.4769	7	1.3806	.7196	-.6639	.8633	-.3807	.1057
15	Winter Flounder	44	.1619	.1396	4	.0000	.0000	16	.4839	.2313	-.4839	.2313	-.0106	.1195
16	Windowpane	36	.0340	.1383	12	.6013	.3476	23	.4572	.1678	.1441	.3860	.0466	.1302
17	Butterfish	81	-.1153	.1983	14	.7529	.3150	17	1.1357	.3645	-.3828	.4818	-.1540	.1833
18	Bluefish	19	.3214	.2665	12	.3226	.1683	7	1.3802	.7634	-1.0576	.7817	.1779	.2522
19	Scup	21	.3177	.2267	11	1.2772	.5525	11	.3258	.1668	.9515	.5771	.4024	.2110
20	Longhorn Sculpin	36	-.0498	.1418	9	.5071	.2606	3	.0000	.0000	.5071	.2606	.0775	.1246
21	Sea Raven	18	.1529	.1905	9	.6577	.4144	16	.4300	.1652	.2277	.4461	.1644	.1752
22	Northern Sea Robin	13	.1998	.2020	12	.6846	.3525	7	.4540	.3677	.2306	.5094	.2040	.1878
23	Amer. Sand Lance	18	-1.2217	.4179	11	.7134	.3018	11	1.6870	.8256	-.9736	.8791	-1.1760	.3774
24	Ocean Pout	14	-.1354	.2390	3	.2310	.2267	5	.6664	.4741	-.4354	.5255	-.1868	.2176
25	Goose Fish	13	-.1728	.2392	11	.2628	.2093	19	.6712	.2661	-.4085	.3386	-.2512	.1954
26	Amer. Lobster	43	-.0764	.1209	15	.5894	.2100	18	.3655	.1704	.2239	.2704	-.0263	.1104
27	Jonah Crab	5	-1.0961	.5920	16	.4642	.2074	27	.8238	.3279	-.3597	.3880	-.5810	.3245
28	Rock Crab	29	-.4385	.1488	19	.6774	.2519	29	.9752	.2992	-.2978	.3911	-.4207	.1391
29	Sea Scallop	19	.3116	.1614	9	.2310	.1331	6	.6931	.4297	-.4621	.4498	.2233	.1520
30	Shortfin Squid	64	-.5004	.1746	10	.5886	.3049	17	1.0700	.4334	-.4814	.5299	-.4985	.1658
31	Longfin Squid	80	-.0105	.1633	16	1.1250	.4741	17	1.3970	.5038	-.2720	.6918	-.0243	.1589
32	Bay and Striped Anchovy Combined	7	1.9778	1.0729	7	3.5186	1.7880	5	5.1205	2.7365	-1.6019	3.2688	1.6297	1.0194

TABLE 3.3

CATCH IN WEIGHT

NO.	SPECIES	BOTH NONZERO			A.NONZERO D.ZERO			A.ZERO D.NONZERO						
		#PAIRS	MEAN	STAND ERROR	#PAIRS	MEAN	STAND ERROR	#PAIRS	MEAN	STAND ERROR	DIFF. MEAN	STAND. ERROR	ESTIMATE	STAND. ERROR
1	Smooth Dogfish	10	.0551	.2425	6	2.4826	1.1533	8	3.4733	1.3641	-.9907	1.7863	.0361	.2403
2	Spiny Dogfish	33	-.3556	.2433	10	2.5624	1.0804	11	2.5255	.8603	.0369	1.3811	-.3438	.2396
3	Winter Skate	43	-.1559	.1914	8	2.0419	.8819	11	2.5011	.9447	-.4592	1.2923	-.1624	.1893
4	Little Skate	55	-.1143	.1079	21	.7703	.2853	17	1.1513	.3756	-.3810	.4717	-.1276	.1052
5	Silver Hake	95	.1723	.0747	16	.3066	.1557	13	1.2148	.4570	-.9083	.4828	.1471	.0739
6	Atlantic Cod	12	-.5914	.4649	5	3.1104	1.5477	3	3.9963	2.3170	-.8859	2.7864	-.5994	.4585
7	Haddock	10	-.2444	.3535	3	2.6654	1.8353	6	2.1186	1.0568	.5469	2.1178	-.2230	.3487
8	White Hake	8	.3374	.3499	12	2.0549	.6454	12	~~2~~.0173	.6742	.0376	.9333	.3004	.3277
9	Red Hake	36	.3582	.1858	13	.6349	.2846	19	.9625	.3310	-.3276	.4365	.2530	.1709
10	Spotted Hake	16	-.1491	.1382	8	.9876	.4176	8	.9450	.4475	.0426	.6121	-.1398	.1348
11	American Plaice	5	-.5414	.6859	5	1.6953	.9776	1	2.7081	.0000	-1.0128	.9776	-.6969	.5615
12	Summer Flounder	29	.0713	.1695	11	2.8378	.8842	5	2.6152	1.1661	.2226	1.4634	.0733	.1684
13	Fourspot Flounder	53	-.0040	.0963	11	1.0732	.4339	16	.7532	.3307	.3200	.5456	.0058	.0949
14	Yellowtail Flounder	43	-.2999	.1251	5	1.4991	.8161	7	1.7449	.6272	-.2458	1.0293	-.2991	.1242
15	Winter Flounder	44	.1346	.1446	4	2.2043	1.0037	16	1.8175	.4821	.3868	1.1135	.1388	.1434
16	Windowpane	36	.1135	.1712	12	.2257	.1623	23	1.1653	.3594	-.9397	.3943	-.0535	.1570
17	Butterfish	81	-.0796	.1286	14	.5895	.2549	17	1.3903	.4612	-.8007	.5269	-.1201	.1249
18	Bluefish	19	.1435	.2661	12	1.9613	.7428	7	3.6719	1.4393	-1.7107	1.6197	.0948	.2626
19	Scup	21	.4261	.2418	11	1.0089	.5148	11	2.4138	.8804	-1.4049	1.0199	.3286	.2353
20	Longhorn Sculpin	36	-.0987	.1329	9	.9644	.3961	3	.5973	.4175	.3671	.5755	-.0751	.1295
21	Sea Raven	18	.2548	.2218	9	1.5180	.5683	16	1.6428	.4948	-.1247	.7536	.2246	.2127
22	Northern Sea Robin	13	.0101	.1931	12	.8202	.3671	7	.7268	.5172	.0934	.6342	.0172	.1847
23	Amer. Sand Lance	18	-.2677	.2119	11	.8443	.4968	11	1.5818	.5765	-.7375	.7610	-.3015	.2042
24	Ocean Pout	14	-.2743	.3833	3	.9635	.7453	5	1.9910	.9972	-1.0276	1.2449	-.3395	.3663
25	Goose Fish	13	.2097	.4206	11	2.1525	.7944	19	2.5994	.7016	-.4469	1.0599	.1204	.3910
26	Amer. Lobster	43	-.2799	.1612	15	2.7667	.7012	18	2.0594	.5299	.7072	.8789	-.2477	.1585
27	Jonah Crab	5	-1.0757	.7001	16	.8044	.2726	27	.9936	.2623	-.1891	.3783	-.3895	.3328
28	Rock Crab	29	-.1062	.1235	19	.5769	.2468	29	.6782	.1897	-.1013	.3113	-.1055	.1148
29	Sea Scallop	19	.1041	.1514	9	.0770	.0770	6	.5669	.3425	-.4898	.3511	.0109	.1391
30	Shortfin Squid	64	-.2813	.1735	10	1.2309	.4524	17	1.5985	.4684	-.3676	.6512	-.2870	.1677
31	Longfin Squid	80	-.0042	.1178	16	1.1352	.4596	17	1.1851	.3802	-.0499	.5965	-.0059	.1156
32	Bay and Striped Anchovy Combined	7	1.3880	.8358	7	1.5357	.9191	5	2.0447	1.1973	-.5090	1.5094	.9428	.7312

4. A CRYSTAL CUBE FOR COASTAL AND ESTUARINE DEGRADATION

Environmental regulators and decision-makers would like to have a crystal ball that could predict how ecosystems would respond to factors such as stress, pollution or over-fishing. In this way, information on important parameters such as amounts of contaminants entering an estuary, their effect on the biota, the propagation of these effects through the ecosystem, and subsequent recovery after the removal of these stresses, could all be properly considered in the use and protection of important natural resources. In the real world, however, such predictions cannot be made with certainty.

This conceptual crystal cube has a series of faces, each of which represents a specific parameter that can be directly related to marine environmental degradation. At present, ten indices or faces of the cube are being tested: dietary risks from contaminants in marine foods; contaminant stress in sediments; contaminant stress in the water column; human pathogen risks; benthic species and composition; fish and shellfish diseases; reproduction in fish and shellfish; mortality of eggs and larvae of fish and shellfish; reproductive success in marine birds; and oxygen depletion. For details, see Boswell and Patil (1985, 1986), Patil (1984a,b), Patil and Taillie (1985), and Pugh, Patil, and Boswell (1986).

The emphasis in testing these indices is on standardizing long term data sets in order to construct a single summary variable—termed an index—to represent each. This index is based on a variable that measures contamination or, ideally, contamination effects. The choice of such a variable is not easy and usuallly involves extensive data analysis. To be useful, the index--summarizing data--must be sensitive to contamination and relatively insensitive to other factors.

The use of this variable index in the crystal cube analyses is in the separation of "concern or alarm" from "no concern" conditions. Here, concern or alarm does not necessarily mean only the violation of legislated or regulated standards, but indicates that the scientific community is not able to certify that issues of widespread public concern will not arise from existing environmental stress.

The index is calibrated so that when the number falls in the range of 0 to 1, there is "no cause for concern."

A flag is raised as soon as the index becomes 1. The range from 1 to 10 indicates "warning;" something is happening and should be investigated. A value of the index in the range from 1 to 10 indicates that the environment has been adversely affected.

The range above 10 indicates "cause for alarm." The index is designed to be 10 when there is scientific reason for grave ecological concern.

The fundamental concept underlying the use of a single summary variable for each of the ten measures of environmental degradation is to compare conditions in a stressed estuary or coastal area with those from a clean region. The crystal cube with ten faces, each with a single summary variable representing one important environmental pollution parameter, will flag cases where legal or scientific benchmarks for concern or alarm are exceeded.

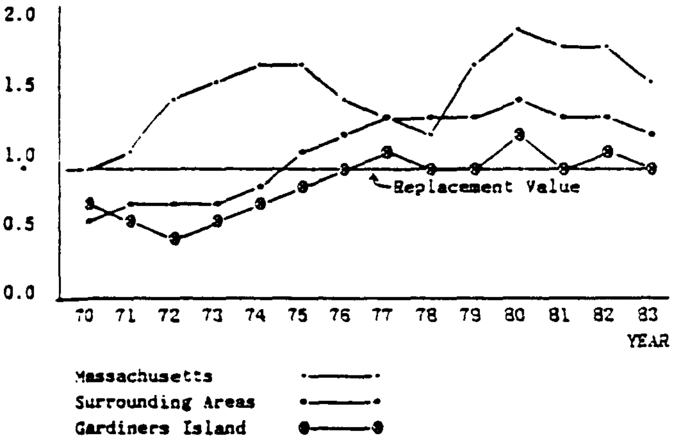
Ultimately, this technique can assist the environmental manager or regulator, who must evaluate large data sets, to narrow attention to those specific environmental parameters where there are serious problems. The crystal cube is not intended to be the "ideal" crystal ball desired by environmental managers, but it does provide a potential framework for evaluating and comparing different environmental measures that must ultimately be weighed not only against each other, but also against other (e.g., economic, aesthetic, etc.) considerations. Thus, the crystal cube could develop into a valuable tool to help define or delineate "unreasonable degradation" and make environmental decision-making a more systematic science, reflecting an integration of the complexity in ecosystems.

5. REPRODUCTIVE SUCCESS OF MARINE BIRDS ON THE EAST COAST

Ospreys are large marine birds nesting and fishing on the east coast of the United States. They nest in accessible areas using the same nesting sites from year to year, and they are tolerant of humans. This has permitted the entire osprey population to be censused at regular intervals since 1974 (see Spitzer, Poole, and Scheibel 1983). In common with other shore birds, reproductive potential of the osprey is sensitive to the presence in the environment of toxicants such as DDT. At the same time, osprey reproduction is much affected by naturally occurring stresses such as wind, storms and food shortage. Thus, any index of anthropogenic impact upon osprey reproduction must carefully incorporate the effects of natural variation.

Figure 5.1 gives three-year moving-average plots of the reproductive success of osprey at several locations along the East Coast. Here, reproductive success is measured as the average number of young fledglings in the active nests. It is evident that there is both spatial and temporal variablity. Much of the temporal variability is attributed to the effects of DDT. In the early 1970's there was extensive DDT pollution which gradually cleared from the environment after its use was banned. The years 1973 to 1979 are transition years when the effects of DDT were still present. The 1980's appear to be essentially free from DDT effects upon osprey reproduction.

Figure 5.1 AVERAGE OSPREY YOUNG PER ACTIVE NEST
(3 year running averages)



For the reasons just described, the five years from 1980 to 1984 were selected as the reference or control period used to calculate a reference value, R, using the effects of natural variability in the environment. The index, in its basic form, is given by

$$I = R/Y, \qquad (5.1)$$

where R is the reference value for reproductive success, expressed in young per active nest, and Y is the reproductive success observed in the year and the region being indexed. The reference value R is the estimated 10th percentile of the distribution of reproductive success during the unstressed reference period from 1980 to 1984. Thus, index values greater than 1 indicate that reproductive success is so low that it could occur only one year in 10 under unstressed conditions.

The index is constructed to flag cases where the reproductive success falls short of the reference values.

The index is calibrated in the range 0 to 1 using data from 1980 to 1984. On the other hand, expert opinion of knowledgeable biologists is necessary to calibrate the index in alarming situations. When the reproductive rate of the ospreys drops below about .8 young per active nest, then the population tends to decrease (Spitzer 1985). The basic index, using a reference value of R=1.7 young per active nest calculated from the combined data from the East Coast, was calibrated to take the value 10 when the reproductive success is .8 young per active nest. The index then becomes

$$I = (R/Y)^c = (1.7/Y)^3 \qquad (5.2)$$

where c is the constant used for calibration.

The values of the index (5.2) are tabulated below for the four regions (i) the Northeast Coast from New York City to Boston, (ii) Massachusetts, (iii) Gardiners Island located off the tip of Long Island, and (iv) area surrounding New York City.

Index of Osprey Reproduction

Region	1969	1970	1971	1972	1973
Northeast Coast	35.0	22.8	19.7	28.0	11.2
Massachusetts	26.6	4.4	11.7	4.4	.6
Gardiners Island	17.9	17.9	35.0	1643.6	29.6
New York City	66.2	61.7	20.7	29.6	66.2

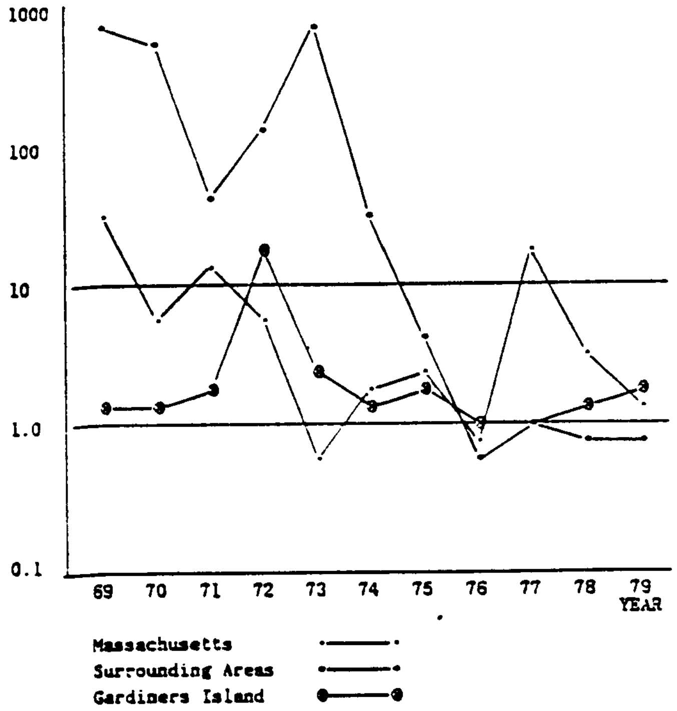
Region	1974	1975	1976	1977	1978	1979
Northeast Coast	7.2	7.2	2.0	4.6	4.2	3.6
Massachusetts	1.7	1.9	.8	18.8	3.7	1.5
Gardiners Island	11.7	22.8	5.7	4.1	10.0	21.7
New York City	17.1	6.1	2.5	3.2	2.9	2.9

The values in the above table were calculated using the global reference value R = 1.7 obtained from the combined data for the entire Northeast Coast. The population of ospreys on Gardiners Island is stressed by a limited food supply. Using a global reference value for such populations results in index values perpetually in the warning or alarm range. A local index based upon a local reference value is to be preferred in such cases because we are attempting to index stresses of authropogenic origin. For ospreys a local index was calculated for each region using the years 1980 to 1984 as a reference period. The resulting index is shown in Figure 5.2 for each of the three local regions. Notice, in particular, that the local Gardiners Island index approaches the value 1 as the DDT passes from the environment. This was not the case when a global reference value was used.

6. LOW DISSOLVED OXYGEN: AN INDEX

When the amount of oxygen dissolved in the water drops below 5 mg/l the ecosystem becomes stressed. The biological response is species dependent and varies with the temperature, the concentration level as well as the duration of exposure. Three responses have been identified for incorporation into an index: mortality (10% of surf clams), avoidance (by 50% of red hake), and reduced growth (15% for winter flounder over the summer season). These species were chosen as the most sensitive from among important species for which data were available.

Figure 5.2 INDEX I_1 FOR OSPREY DATA



To calculate an index value, dissolved oxygen data must be available for a given location on a daily basis throughout the summer season. Low-dissolved-oxygen (low DO) episodes, defined to occur when the dissolved oxygen concentration drops below 5 mg/l, occur in the summer months. An index value for the season is the maximum of index values calculated for the low DO episodes throughout the summer. The value calculated for a given low DO episode is, in turn, the sum of three values corresponding to the three responses. To calibrate the index (see discussion in Sections 4 and 5) to be 10 in an alarming situation the value corresponding to mortality is multiplied by 10 before adding.

Before an index can be calculated, it is necessary to know the intensity of each response to low DO concentrations; this is estimated by laboratory studies. The resulting dose-response curves provide reference values to compare with the observed low DO concentrations. Three curves, giving the day's exposure to produce the indexed response as a function of the DO concentration must be determined. Since the effect of low DO varies with temperatures, the average summertime temperature in the region to be indexed is used.

Let c_i be the observed DO concentration on the *ith* day of a low DO episode, $i=1,2,\ldots,n$. Let m_i, a_i and r_i stand for the days exposure to a DO concentration of c_i to produce the mortality, avoidance and reduced growth respectively. The index can be formulated as

$$I = 10\sum_{i=1}^{n}\frac{1}{m_i} + \sum_{i=1}^{n}\frac{1}{a_i} + \sum_{i=1}^{n}\frac{1}{r_i}$$

$$= \sum_{i=1}^{n}\left[\frac{10}{m_i} + \frac{1}{a_i} + \frac{1}{r_i}\right] = \sum_{i=1}^{n} q_i ,$$

where q_i is the combination of the three exposure curves. The data is summarized into DO categories and the above calculation is simplified by using the frequencies. The index becomes

$$I = \sum_{i=1}^{k} f_i q_i \qquad (6.1)$$

where f_i is the number of days that the low DO episode has concentrations in the *ith* category (corresponding to the value q_i) and where k is the number of categories. The calculation of an index value is illustrated in Table 6.1.

TABLE 6.1

Calculation of the dissolved oxygen index for a low DO episode at the Narrows*, New York Harbor, Summer 1975

DISSOLVED OXYGEN CONCENTRATION	FREQUENCY OBSERVED AT DO CONCENTRATION	COMBINED RESPONSE CURVE	CONTRIBUTION TO THE INDEX
1.0	0	-	0.0
1.2	0	4.55	0.0
1.4	2	4.35	8.7
1.6	0	4.17	0.0
1.8	1	4.08	4.1
2.0	3	3.91	11.7
2.2	1	3.70	3.7
2.4	2	3.55	7.1
2.6	2	3.33	6.7
2.8	2	3.11	6.2
3.0	1	2.86	2.9
3.2	2	2.63	5.7
3.4	1	2.38	2.4
3.6	3	2.08	6.2
3.8	2	1.79	3.6
4.0	2	1.52	3.0
4.2	3	1.11	3.3
4.4	4	0.67	2.7
4.6	3	0.22	0.7
4.8	1	0.16	0.2
5.0	2	0.01	0.0
		TOTAL INDEX	78.9

*The Narrows is known to be polluted.

This index has undergone many changes from the original formulation and is currently being looked at. The final form of the index has not been fixed at this time. The example given in Table 6.1 is for illustration purposes; the data is for a summer season which may not correspond to one low-dissolved-oxygen episode required for the purposes of this index.

7. COMBINING BIO-ASSAY RESULTS FOR EXTRAPOLATION OF CHRONIC EFFECT THRESHOLDS FOR RISK ASSESSMENT

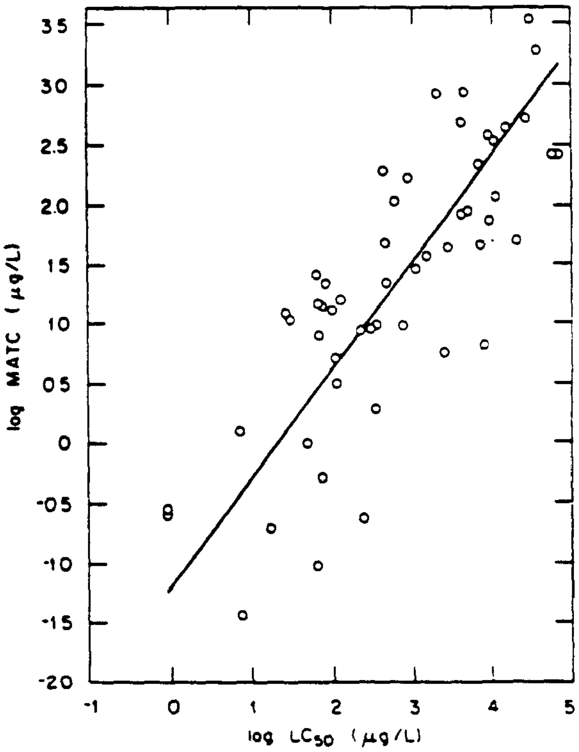
7.1 Introduction

Ecological effects of toxic chemicals are commonly assessed by estimating a "safe" exposure level, below which no effects will occur. To protect organisms at their most sensitive stages, life cycle tests or, in some cases, early life stage tests are necessary for estimation of chronic effect threshold levels. It is not feasible to conduct tests for every possible toxicant and species of interest. Instead, "safe" levels are commonly extrapolated from laboratory test results of a few standard test species and particular life stages by applying correction factors and subjective judgement. Suter, et al. (1986a) propose a more structured approach called, "Analysis of Extrapolation Error" (AEE) of extrapolating chronic effect thresholds. Its main features and advantages over traditional methods lie in the explicit quantification of the consequence of exceeding the estimated safe level, of interspecies differences in sensitivity between tested and extrapolated species, and the variable relationship between acute and chronic effects of chemicals. See also Linder, Patil, Suter, and Taillie (1986).

7.2 Acute-Chronic Extrapolation

AEE is based on statistical analysis of acute and chronic toxicity test data sets collected using uniform experimental protocols. For each species and chemical pair, two different studies are conducted to determine long-term low-level effect, or maximal allowable toxic concentration (MATC) and 96-hour LC50 high-level effect to achieve 50% mortality in 96 hours. If enough of these results are available, a functional relationship, the so called acute-chronic extrapolation can be estimated. It is used in turn to extrapolate from the LC50 to the MATC (see Figure 7.1). Note that each LC50-MATC pair is the result of a different study reported in the toxicological literature. To ensure compatability, only results obtained under similar experimental conditions with uniform protocols are used for establishing the acute-chronic relationship.

Estimated error (residuals) about that relationship and error of the parameter estimates determine the error of an extrapolated MATC, similar to prediction variances in ordinary linear regression analysis. This allows the calculation of the risk that a given environmental concentration of the chemical being assessed exceeds the extrapolated MATC for the species of interest.



Logarithms of MATC values from life-cycle or partial life-cycle tests plotted against logarithms of 96-h LC_{50} values determined for the same species and chemical in the same laboratory. The line is derived by an errors-in-variables regression;

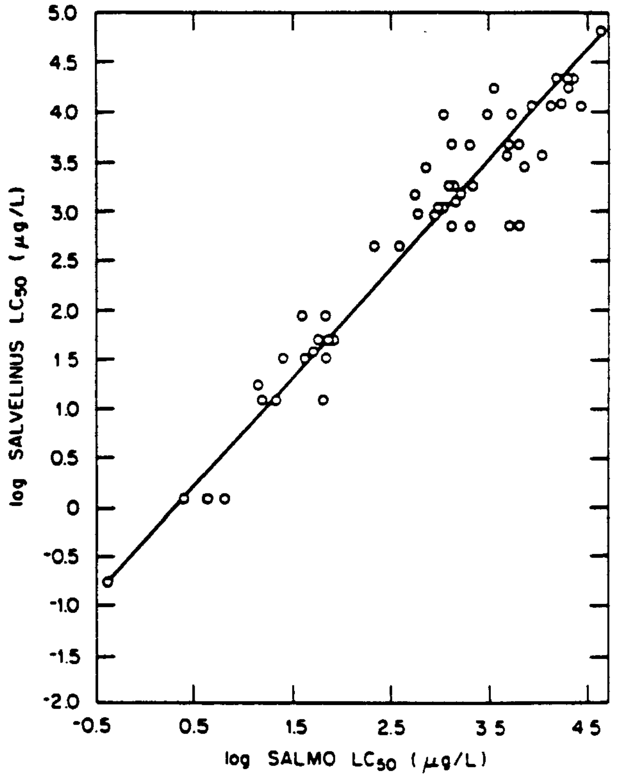
Figure 7.1 (from Suter et al., 1986a)

Suppose we want to assess the chronic effect of a given chemical on species A of fish. If an LC50 is available for this species-toxicant pair, the acute-chronic extrapolation can be applied directly. The variability of the data about any fitted curve is quite large. As a result, different classes of curves might be considered for defining the extrapolation relationship. In the applications examined so far, logarithmic transformation in both variables produced a linear trend and homogeneity (equal variances) about that trend.

7.3 Taxonomic and 2-step extrapolation

In the case where no test results on species A of interest are available, another test species B has to be chosen for the purpose of extrapolation.

The uncertainty due to extrapolating from B to A depends on the difference of the sensitivities of species A and species B to the chemical. This difference is assumed to be proportional to the "taxonomic distance" between A and B. For this reason, extrapolation relationships are estimated between taxa having the next higher taxon in common. This is done by pairing LC50's of common chemicals to members of the two taxa, whenever there are enough such pairs to allow curve fitting. The resulting curve is used for extrapolating the LC50 of the species of interest from the LC50 of the test species. This is called taxonomic extrapolation. Figure 7.2 depicts an example of a taxonomic extrapolation between two different genera that are members of the family Salmonidae. The LC50's data base is compiled either from a single laboratory or from several laboratories (Suter et al. 1986b). The data base is screened to insure compatability with respect to testing conditions. As with the acute-chronic extrapolation, the data are log transformed to produce linearity and homogeneity.



Logarithms of LC$_{50}$ values for Salvelinus plotted against Salmo. The line is determined by an errors-in-variables regression;

Figure 7.2 (from Suter et al., 1986b)

Taxonomic and subsequent acute-chronic extrapolation are combined in order to extrapolate to a chronic effect threshold. Thus, Z = c+dY = c+d(a+bX), results from combining Y = a+bX, the estimated line for taxonomic extrapolation and Z = c+dY, the estimated line for acute-chronic extrapolation. The variance of an extrapolated MATC is calculated under the assumption of statistical independence between the set of variables associated with the two extrapolations. Estimated variance of an extrapolated MATC is quite large especially when the extrapolation requires more than one step. Assuming a normal distribution for the extrapolated MATC (Suter et at. 1986a) is therefore not likely to affect the resulting risk calculation too strongly.

7.4 The Data

We focus in the following on the acute-chronic extrapolation. Most of the problems and issues discussed arise equally in the context of the taxonomic extrapolations. Let (X_i,Y_i) be an LC50-MATC pair for a particular toxicant species combination (i = 1,...,n). The following are the features of the acute-chronic "data set" of Fig. 7.1:

(i) Each point (or pair) represents a reported result from a bio-assay experiment.

(ii) Different points result from different studies.

(iii) The collection of points has been gathered from the literature. Hence the (X_i,Y_i) may not constitute a random sample from the population of all possible LC50-MATC pairs.

(iv) Since (X_i,Y_i) are estimates of threshold concentrations, they are themselves random quantities. There is considerable uncertainty about their "true" values.

7.5 Combining Estimates

Traditional methods of extrapolation are often based on the use of a single test species such as fathead minnow for fresh water fish. The ratio of its MATC to LC50 is multiplied by X_0, the LC50 for the species of interest; using the above notation, this is

$$\hat{Y} = X_0 \cdot Y_j/X_j, \text{ where } (X_j,Y_j)$$
denotes (LC50, MATC) of the test species.

This provides a point estimate of the chronic threshold concentration. It is subsequently "scaled down" by correction factors accounting for the uncertainties due to extrapolation and natural variabilities. The final value for the extrapolated MATC depends strongly on the test species chosen as well as on the particular sources of uncertainties considered.

An improved combined estimate can be obtained by using the test results of several toxicants and several species. Let $b_i = Y_i/X_i$ be the "slope" from the test on the i*th* toxicant-species pair. A combination estimate can be formed by means of a linear combination of the individual estimates using weights w_i:

$$b = \sum_{i=1}^{n} w_i b_i .$$

In some cases, when the weights are chosen proportional to the statistical influence of the corresponding observations, the resulting combination estimate turns out to correspond to a particular estimate of that curve. For linear regression through the origin, the slope estimate is obtained by choosing the weights proportional to the squared lengths of the X_i:

$$b = \frac{\Sigma\, x_i y_i}{\Sigma\, x_i^2} = \frac{\Sigma\, x_i^2\, y_i/x_i}{\Sigma\, x_i^2} = \Sigma\, w_i b_i ,$$

$$\text{where } b_i = y_i/x_i ,$$

$$\text{and } w_i = \frac{x_i^2}{\Sigma\, x_i^2} ,$$

$$(i=1,\ldots,n) .$$

Thus, combining extrapolation factors (estimates) b_i is equivalent to estimating an extrapolation curve for the collection of points (X_i, Y_i). This motivates the procedures described below. Notice that we have not considered any of the technicalities such as variable transformation or correction for the intercept.

The main advantage of estimating an extrapolation curve over traditional method lies in the availability of standard error estimates. Thus the method provides explicit quantification of the sources of errors involved. Combining all available LC50-MATC pairs reduces the uncertainty by using the largest possible data set. On the other hand, it increases the variability because a wide variety of toxicants and species are "lumped" together. Extrapolation by using results of one particular class of chemicals reduces this variability. However the representativeness of the remaining species might be in question after the data have been partitioned by chemical class. The use of the appropriate data set for a given extrapolation problem needs to be carefully examined.

7.6 The Model

Logarithmic transformation of both X_i and Y_i produces linearity and homogeneity to a satisfactory degree given the natural variabilities. We propose an errors-in-variables (EIV) model for estimation of the extrapolation line for the data described in 7.4. The assumptions underlying ordinary least-squares (OLS) regression analysis are clearly violated. In the errors-in-variables model (X_i, Y_i) are assumed to have been recorded with error. They represent unknown mathematical quantities (U_i, V_i). Linearity is assumed between the U_i and V_i, resulting in the model:

$$X_i = U_i + \delta_i \qquad V_i = \alpha + \beta U_i, \quad i = 1, \ldots, n.$$
$$Y_i = V_i + \varepsilon_i$$

Normal distributions with zero means are commonly assumed for the errors δ_i, ε_i, with

$\mathrm{var}(\delta_i) = \sigma_\delta^2$, $\mathrm{var}(\varepsilon_i) = \sigma_\varepsilon^2$, $\mathrm{cov}(\delta_i, \varepsilon_i) = 0$, and $(\delta_i, \varepsilon_i)$ independent of $(\delta_j, \varepsilon_j)$ for $i \neq j$.

Two EIV models have been studied extensively: (Kendall and Stuart, 1979; Gleser 1983).

(i) The structural model: the U_i are assumed to be a random sample.

(ii) The functional model: no assumption on the U_i.

For identifiability in both models, one of the variance parameters has to be assumed known a priori. In the classical case, this is $\lambda = \sigma_\varepsilon^2/\sigma_\delta^2$.

More complicated models would be more realistic for the extrapolation problem. These are:

(iii) The ultrastructual model: U_i random but with different locations for different i (Dolby, 1976).

(iv) The model with error in the equation (Schneeweiss, 1976).

Both (iii) and (iv) can not be distinguished from the functional model (ii) unless replicate observations are available or additional a priori assumptions about the error structure are made (Gleser, 1985).

Maximum likelihood estimators of the slope β and the intercept α for both, the structural and the functional model, are:

$$\hat{\beta} = h + \mathrm{sign}(SXY)(h^2+\lambda)^{1/2};$$

$$\text{where } h = \frac{SYY-\lambda SXX}{2SXY}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X};$$

where bars denote averages and SXX, SXY, SYY the usual corrected sums of squares.

The corresponding OLS slope $b = SXY/SXX$ is smaller than $\hat{\beta}$ in absolute value, thus downward biased. $\hat{\beta}$ can also be obtained by minimizing the sum of squared distances from (X_i, Y_i) to the line under a vertical angle with tangent of $\hat{\beta}/\lambda$ (Mandel, 1984). Such a least-squares interpretation of $\hat{\beta}$ allows for straightforward generalizations to weighted EIV in the case of unequal variances (Sprent, 1966). Thus reporting biases and inhomogeneities resulting from combining the LC50-MATC estimates can be incorporated in the model. This produces the estimates:

$$\hat{\beta} = h^W + (h^W+\lambda)^{1/2}; \text{ where } h^W = \frac{SYY^W - \lambda SXX^W}{2SXY^W}$$

$$\hat{\alpha} = \bar{Y}^W - \hat{\beta}\bar{X}^W; \quad \bar{X}^W = \frac{\Sigma\, w_i X_i}{\Sigma\, w_i}$$

$$SXX^W = \Sigma\, w_i (X_i - \bar{X}^W)^2, \text{ etc.}$$

7.7 Risk Calculation and Evaluation

The final risk calculation is determined by the statistical distribution of the MATC that has been extrapolated from an input LC50. Exact distributions of the EIV estimators are not tractable. Given the large variabilities, estimates of asymptotic standard error combined with normality assumption are considered accurate enough in this context. For future refinements, resampling methods such as the bootstrap might be applied to obtain small sample distributions or standard error estimates.

For the approximate variance of an extrapolated MATC, we add a variance term $(\hat{\sigma}_\varepsilon^2)$ to the variance of a fitted Y-value as given in Mandel, 1984. This results in the following formula for the weighted EIV model:

$$\hat{\mathrm{var}}(Y_{new})$$

$$= \hat{\sigma}_\varepsilon^2 + \hat{\beta}^2\hat{\sigma}_\delta^2 + S_\varepsilon^2\left[\frac{1}{\Sigma w_i} + \frac{(1+\hat{\beta}^2/\lambda)^2(x_0-\bar{X}^W)^2}{SXX^W + 2\hat{\beta}SXY^W/\lambda + \beta^2 SYY^W/\lambda^2}\right]$$

$$\text{where } S_e^2 = \frac{\hat{\beta}^2 SXX^W - 2\hat{\beta}SXY^W + SYY^W}{n-2}$$

$$\hat{\sigma}_\varepsilon^2 = \frac{S_e^2}{1+\hat{\beta}^2/\lambda}, \quad \hat{\sigma}_\delta^2 = \hat{\sigma}_\varepsilon^2/\lambda.$$

It has been frequently suggested (Lindley, 1946; Kendall and Stuart, 1979) to use OLS regression if the purpose of the analysis is prediction even when the X's have been measured with error. Prediction, or in general regression analysis is based on the covariation of the two random variables X and Y. Predictions by means of the conditional properties of Y given X are possible for the structural EIV model. The situation is different in the functional model. Strictly speaking, the (X_i, Y_i) constitute different random variables for $i = 1, \ldots, n$. The only existing relationship lies in the proposed structure (in this case, the line) for the location of the means. Thus this is not a classical prediction problem, since no conditional means or variances are involved. It is for this reason, we propose to use EIV estimates for the extrapolation problem.

The method of combining estimates for extapolating MATC's is based on purely statistical grounds. Relative magnitudes of estimated standard errors seem to be appropriate from a biological point of view. For the applicability of the methods, the results (absolute magnitudes) have to be evaluated in terms of their biological meaning. This has been done extensively (Suter et al., 1986a) by comparing extrapolated MATC's with measured MATC's for toxicant-species combinations, where the results are available. In addition the method was compared to the traditional approaches using only results on fathead minnow. It has been generally found to out perform the old methods.

8. COMBINING CONCLUSIONS ACROSS SPECIES

The NOAA Chesapeake Bay Stock Assessment Committee has been set up to help develop a plan to establish a cooperative stock assessment program. The Center for Statistical Ecology and Environmental Statistics has been studying various continuous and categorical multiple time series methods to partition the effects of fishing mortality, natural mortality and the effects of pollutant loadings on stock sizes (Boswell, Linder, Ord, Patil, and Taillie, 1986). The data set used to evaluate these methods was compiled from historical environmental data, pollution data, and fishing data (see Summers et al. 1984). As much as possible, the environmental and pollution variables were chosen to be meaningful in terms of the fish stocks to be investigated. However, the pollution variables are macro-pollution variables which give gross indications of the corresponding pollution loadings.

Examples of environmental variables are average monthly air temperature, river temperature and flow, wind speed and direction, etc. Examples of pollution variables of regions in or near the selected water systems are human population size, employees in manufacturing industries, sewage volume discharge, acreage in improved farmland, total annual volume dredged, five-day biochemical oxygen demand for loadings from treatment plants, minimum 28 day average summertime dissolved oxygen, etc. The fish data, consisting of information from various sources, was combined to give a stock index in the form of catch per unit effort for the dominant fishing gear used in the region. Different species were chosen for different parts of the Chesapeake Bay. For the Potomac River system, the species chosen are striped bass, American shad, American oyster and blue crab.

The results of a study of the pollution effects on fish stocks would be of interest to managers with the job of deciding what pollution is in need of abatement programs. It may turn out that different fish stocks are affected differently by the pollutants. If the study could provide some measure of impact for each pollutant on each species and if the manager can provide weights giving the importance of each species, then the problem can be approached by the method of combining expert opinions, described in the next section. With macropollution variables such as those included in the study, clear-cut results are not to be expected. This study was mainly to identify, adapt, and develop the statistical techniques as needed.

The first technique considered is that of multivariate time-series regression. This requires the assumption of some functional form to incorporate the effect of the variables on the stock size variable. The usual assumption of a linear model was used and all biologically meaningful lags were incorporated.

Categorical regression was used by Summers et al (1984) and was studied here as a starting point for modifications to incorporate meaningful biological concepts into the lag structure. The methods tried incorporate a distributed lag structure and a combination of continuous and categorical methods.

Transfer function modeling was also tried.

It is possible that various methods would yield different results. If the correct method to use is unknown, then the combining of the results of various studies using different methods may provide better results than any one study by itself. This problem is analogous to the problem of combining expert opinions as outlined in the next section.

9. COMBINING CONCLUSIONS FROM EXPERTS

The necessity for combining probabilities could arise in two situations—described respectively as the group decision problem and the panel of experts problem. In the first, individuals with different probability judgements (and different preferences) have to make a joint decision. Under certain circumstances, this joint decision could be the result of maximizing a group expected utility where the expectation is taken with respect to a combined or group probability distribution. In general, the group decision problem is intractable. (References to various aspects of this problem include Arrow (1951), Hyland and Zeckhause (1980) and Wilson (1968).)

The second situation, the panel of experts problem, is set in the context of a single decision maker who wishes to combine <u>information</u> obtained from various experts, rather than opinions. While, in the group decision problem, agreement on probabilities would facilitate solution of the problem, divergence of information is beneficial for a decision maker seeking independent sources of information as input into his or her judgement.

Morris (1977), Winkler (1981), Lindley (1983,1985) and others produce resolutions of the panel of experts problem. In Lindley (1983), the decision maker has a diffuse prior on the quantity of interest. The conditional distribution of expert assessments given the true value of the parameter is multivariate normal with the individual expert's assessment of the mean allowing for bias and with the decision maker having to specify the covariances between the different expert assessments. The decision maker's posterior mean, given the expert assessments, is then shown by Lindley to be a weighted average of the experts' assessed means.

Note that it is always at least as good to include additional experts (if they are free) as not to include them. This is a different way of stating the standard result on the non-negative value of information. However, this assumes that all experts have appropriate incentives to gather and to report information.

In certain circumstances, the likelihood function of expert assessments may be impossible to specify. In this case, insights obtained from group decision theory and axiomatic approaches to combination may help in aggregating assessments without explicitly calculating posterior distributions. Axiomatic approaches are to be found in Madansky (1964), Morris (1983) and the extensive literature in Section 3 of Genest and Zidek (1986). Madansky (1964) considered the linear opinion pool, a weighted arithmetic average of probabilities, and showed that such a linear opinion pool was not "externally Bayesian with a fixed constitution." That is, it was impossible to find a set of non-negative weighting constants such that a posterior based on a common likelihood and a weighted prior would be the same as a weighted posterior based on a common likelihood and different priors. Raiffa (1968) argues that, in this case, the priors should be combined.

The Wilson (1968) theory of syndicates yields a geometric average of individual probability distributions (or an arithmetic average of log odds ratios) as the appropriate combined distribution. This avoids the difficulty found by Madansky.

An example due to John Pratt (cited in Raiffa

(1968)) shows that linear opinion pools have an additional deficiency of not preserving the independence of events after combination. Genest and Zidek (1986) discuss recent investigations of this independence preservation property.

Another reaction to the lack of a well-specified likelihood function is to dispense with probability theory and rely on the Dempster-Shafer theory of combining evidence. (Shafer (1976), Krantz and Miyamoto (1983).)

ACKNOWLEDGEMENTS

The research effort leading to this paper and presentation has been partially supported by research grants of the National Oceanic and Atmospheric Administration under the auspices of the Chesapeake Bay Stock Assessment Committee, the Northeast Fisheries Center, Woods Hole, Massachusetts and the Ocean Assessments Division, Rockville, Maryland to the Center for Statistical Ecology and Environmental Statistics of the Department of Statistics at the Pennsylvania State University, University Park. The co-authors of this paper are members of the Center; G. P. Patil is the Director. During this year, he is also a Visiting Professor of Biostatistics at the Harvard School of Public Health.

REFERENCES

1. Anandalingam, G. and Chatterjee, K. (1986). Personal communication.
2. Arrow, K. J. (1951). *Social Choice and Individual Values.* Yale University Press.
3. Babu, G. J. (1986). A note on comparison of conditional means. Preprint.
4. Babu, G. J., Pennington, M., and Patil, G. P. (1986). Estimation of relative fishing power of different vessels. In *Oceans 86 Proceedings: Vol. 3: Monitoring Strategies Symposium*, pp. 914-917. Washington, D.C.
5. Boswell, M. T., Linder, E., Ord, J. K., Patil, G. P., and Taillie, C. (1986). Time series regression methods for the evaluation of the causes of fluctuation in fishery stock sizes. In *Oceans 86 Proceedings: Vol. 3: Monitoring Strategies Symposium*, pp. 940-945. Washington, D.C.
6. Boswell, M. T. and Patil, G. P. (1985). Marine Degradation and Indices for Coastal and Estuarine Monitoring and Management. A research paper presented at the spring meetings of the American Statistical Association and the Biometric Society, ENAR, North Carolina State University, Raleigh, N.C.
7. Boswell, M. T., and Patil, G. P. (1986). Field based coastal and estuarine statistical indices of marine degradation. In *Oceans 86 Proceedings: Vol. 3: Monitoring Strategies Symposium*, pp. 929-933. Washington, D.C.
8. Byrne, C. J., and Fogarty, M. J. (1985). Comparison of fishing power of two fisheries research vessels. Preprint.
9. Dolby, G. R. (1976). The ultrastructural relation: A synthesis of the functional and structural relations. *Biometrika*, 63, 39-50.
10. Genest, C., and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1), 114-148.
11. Gleser, L. J. (1983). Functional, structural and ultrastructural errors-in-variables models. *Proc. Bus. Econ. Statist. Sect.*, pp. 57-66, Washington, D.C.: American Statistical Association.
12. Gleser, L. J. (1985). A note on G. R. Dolby's unreplicated ultrastructural model. *Biometrika*, 72, 117-124.
13. Hennemuth, R. C., Palmer, J. E., and Brown, B. E. (1980). A statistical description of recruitment in eighteen selected fish stocks. *J. Northwest Atlantic Fishery Science*, 1, 101-111.
14. Hennemuth, R. C. and Patil, G. P. (1983). Implementing statistical ecology initiatives to cope with global resource impacts. In *Proc. of International Conference: Renewable Resource Inventories for Monitoring Changes and Trends.* J. F. Bell and T. Atterbury, eds. Corvallis, Oregon, pp. 374-378.
15. Hennemuth, R. C., Patil, G. P., and Ross, N. P. (1986). Encountered data analysis and interpretation in ecological and environmental work: Opening remarks. In *Oceans 86 Proceedings: Vol. 3: Monitoring Strategies Symposium.* Washington, D.C.
16. Hennemuth, R. C., Patil, G. P., and Simberloff, D. (1986). Advanced Research Conference on Frontiers of Statistical Ecology. *Intecol Newsletter*, 16(1), 4.
17. Hennemuth, R. C., Patil, G. P., and Taillie, C. (1985). Can we design our encounters? CM1985/D:9, International Council for the Exploration of the Sea, London.
18. Hylland, A., and Zeckhauser, R. (1980). The impossibility of Bayesian group decision making with separate aggregation of beliefs and values. Harvard University. (Mimeo).
19. James, W., and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press, pp. 361-379.
20. Kendall, M. G. and Stuart, A. (1979). *The Advanced Theory of Statistics*, Vol. 2, 4*th* edition. Macmillan, New York.
21. Krantz, D. H., and Miyamoto, J. (1983). Priors and likelihood ratios as evidence. *J. Amer. Statist. Assoc.*, 78, 418-423.
22. Linder, E., Patil, G. P., Suter, G. W., II, and Taillie, C. (1986). Effects of toxic pollutants on aquatic resources using statistical models and techniques to extrapolate acute and chronic effects benchmarks. In *Oceans 86 Proceedings: Vol. 3: Monitoring Strategies Symposium*, pp. 960-963. Washington, D.C. press).
23. Lindley, D. V. (1947). Regression lines and the linear functional relationship. *J. Royal Statist. Society, Series B*, Vol. 9, 218-244.
24. Lindley, D. V. (1983). Reconciliation of probability distributions. *Operations Research*, 31, 866-880.
25. Lindley, D. V. (1985). Reconciliation of discrete probability distributions. In *Bayesian Statistics*, Vol. 2, J. M. Bernardo, et al., eds. North Holland, Amsterdam. pp. 375-390.
26. Madansky, Albert (1984). Externally Bayesian groups. Rand Corp. Memorandum MR 4141-PR, December.
27. Mandel, J. (1984). Fitting straight lines when both variables are subject to error. *J. Qual. Technol*, 16: 1-14.
28. Morris, P. A. (1977). Combining expert judgements. *Management Science*, 23, 679-693.
29. Morris, P. A. (1983). An axiomatic approach to expert resolution. *Management Science*, 29(1), 24-32.
30. O'Connor, J. S. and Dewling, R. T. (1986). Indices of marine degradation: Their utility. *Environmental Management.* (In Press).
31. Patil, G. P. (1984a). On constructing a crystal cube for environmental degradation. Opening Technical Remarks at the Workshop on Indices of Marine Degradation: An Overview for Managers. November 15-16, 1984, Washington, D.C.
32. Patil, G. P. (1984b). Some perspectives of statistical ecology and environmental statistics. In *Statistics in Environmental Sciences*, ASTM STP 845, S. M. Gertz and M. D. London, eds. Amer. Soc. Testing and Materials. pp. 3-22.
33. Patil, G. P. (1985). Fishery and forestry management: Preface. *Amer. Statist.*, 39(4), 361-362.
34. Patil, G. P., Rao, C. R., and Zelen, M. (1986). A computerized bibliography of weighted distributions and related weighted methods for statistical analysis and interpretation of

encountered data, observational studies, representativeness issues, and resulting inferences. Center for Statistical Ecology and Environmental Statistics, The Pennsylvania State University. (Under preparation).

35. Patil, G. P., and Taillie, C. (1981). Statistical analysis of recruitment data for eighteen marine fish stocks. Invited Paper Presented at the Annual Meetings of the American Statistical Association, Detroit, MI.

36. Patil, G. P. and Taillie, C. (1985). A Conceptual Development of Quantitative Indices of Marine Degradation for Use in Coastal and Estuarine Monitoring and Management. A research paper presented at the spring meetings of the American Statistical Association and the Biometric Society, ENAR, North Carolina State University, Raleigh, N.C.

37. Pugh, W. L., Patil, G. P., and Boswell, M. T. (1986). The crystal cube for coastal and estuarine degradation. *Sea Technology*, September 1986, p. 33.

38. Raiffa, H. (1968). *Decision Analysis.* Addison-Wesley.

39. Schneeweiss, H. (1976). Consistent estimation of a regression with errors in the variables. *Metrika*, 23, 101-115.

40. Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, New Jersey.

41. Spitzer, P. R. (1985). A private communication.

42. Spitzer, P. R.; Poole, A. F.; and Scheibel, M. (1983). Initial population recovery of breeding ospreys in the region between New York City and Boston. In: *Biology and Management of Bald Eagles and Ospreys.* Editor D. M. Bird. Harpell Press, Ste. Anne de Bellevue, Quebec.

43. Sprent, P. (1966). A generalized least-squares approach to linear functional relationships. (With discussion). *J. Royal Statist. Soc., Series B*, Vol. 28, 278-297.

44. Summers, J. K., Polgar, T. T., Rose, K. A., Cummins, R. A., Ross, R. N. and Heimbuch, D. G. (1984). Assessment of the Relationships among Hydrographic Conditions, Macropollution Histories, and Fish and Shellfish Stock in Major Northeastern Estuaries. Technical Report, Martin Marietta Environmental Systems.

45. Suter, G. W.,II, Rosen A. E., Linder, E. (1986a). Analysis of extrapolation error. In *User's Mannual for Ecological Risk Assessment.* L. W. Barnthouse, and G. Suter, eds. ORNL-6251, Oak Ridge National Laboratory, Oak Ridge, TN.

46. Suter, G. W.,II, and Rosen, A. E. (1986b). Comparative toxicology of marine fishes and crustaceans. ORNL-TM, Oak Ridge National Laboratory, Oak Ridge, TN. (In press).

47. Verner, J.; Pastorok, R.; O'Connor, J.; Severinghaus, W.; Glass, N.; and Swindel, B. (1985). Ecology community structure analysis in the formulation, implementation, and enforcement of law and policy. *Amer. Statist.*, 39(4), Part 2, 393-402.

48. Wertz, W., and Schneider, B. (1979). Statistical density estimation: A bibliography. *International Statistical Review*, 47, 155-175.

49. Wilson, R. (1968). The theory of syndicates. *Econometrica*, 36, 119-132.

50. Winkler, R. L. (1981). Combining probability distributions from dependent information sources. *Management Science*, 27, 479-488.

DISCUSSION

Lloyd L. Lininger, U.S. Environmental Protection Agency

The problems presented in this paper give an idea of the diversity and difficulty of the problesm that are routinely encountered at the U.S. Environmental Protection Agency. I wish to draw attention to the diversity because I think that it is currently not practical to think of developing a methodology for combining studies that is simultaneously applicable to all problems. I think the problems presented also point out the basic reason we must work on the problem of combining results of studies. Namely, we are unable to do "the" experiment that we believe is required to make a decision. We have to use the information we have from other studies that were designed for a different purpose and possibly augment them with further studies to make the decision. This always requires the assumption of some model, possibly with some error structure included.

I am confident that any attempt to apply the methodologies advocated by the previous speakers to these problems would soon expose the difficulties and uncertainties of those procedures. I do believe that those attempts should be made. They would result in systematic approaches to modeling the underlying problems and focus attention in the most appropriate places.

The recruitment data presented is useful for emphasizing several points. When combining studies we must always keep in mind the "question" we wish to address. The authors state that one of the objectives is to estimate a "universal" recruitment distribution. First, note that the solution will be a "distribution" which is a somewhat unusual outcome for an experiment. Details of the analyses are lacking, but I would be interested in the sequential aspect of each of the individual samples, the possible relationships between species and the way the 18 species were selected from all species. Finally, I would ask why a "universal" distribution is desired. Too frequently, one looks for a question that combining data sets might answer, instead of looking for ways to answer a question by combining data sets and developing an appropriate model.

The "crystal cube" problem is also a long standing problem in statistics. How do we reduce a multivariate model to a one dimension model that will serve as an index of the phenomenon in which we are interested? Unfortunately, the terminology gives no help in deriving or understanding how such an index would be derived. The relationships between faces and some "geometrical" concepts would be helpful before this terminology is accepted.

The problems presented in this paper tend to combine studies each of which collected information on the same problem. Problems which are "linked linearly" as in the exposition by Eddy and Wolpert require different modeling approaches.

Finally, any methodology to combine data across studies assumes some underlying model. If extensive effort has gone into developing a model to combine studies, then it is relatively straight forward to do simulation studies to evaluate the characteristics of the model. None of the presentations exploited this useful technique.

APPENDIX A: ASA/EPA Conference on Statistical Issues in Combining Environmental Studies Program

ASA/EPA CONFERENCE ON STATISTICAL ISSUES IN COMBINING ENVIRONMENTAL STUDIES

OMNI SHOREHAM HOTEL
WASHINGTON, D.C.

OCTOBER 1-2, 1986

American Statistical Association
806 Fifteenth Street, N.W.
Washington, D.C. 20005

ORGANIZING COMMITTEE

Chair: Kinley Larntz, Washington State University-Pullman
Dorothy Wellington, Environmental Protection Agency
Walter S. Liggett, National Bureau of Standards
Emanuel Landau, American Public Health Administration

This Conference is the second in a series of research conferences organized by the American Statistical Association, supported by a cooperative agreement between ASA and the Office of Standards and Regulations, under the Assistant Administrator for Policy Planning and Evaluation, Environmental Protection Agency.

PROGRAM

WEDNESDAY, OCTOBER 1

9:00 a.m.
INTRODUCTION

Kinley Larntz, Washington State University
Dorothy Wellington, Environmental Protection Agency

9:10 a.m.
CONFIDENCE PROFILES:
A BAYESIAN METHOD FOR ASSESSING HEALTH TECHNOLOGIES

David Eddy & Robert Wolpert, Duke University

10:40 a.m.
BREAK

10:55 a.m.
DISCUSSANT

David Lane, University of Minnesota/McGill University

11:30 a.m.
COMPUTER DEMONSTRATION OF CONFIDENCE PROFILES METHODOLOGY

12:15 p.m.
LUNCH

1:30 p.m.
META-ANALYSIS AND ENVIRONMENTAL STUDIES

Larry V. Hedges, University of Chicago

3:00 p.m.
BREAK

3:15 p.m.
DISCUSSANTS

Chao Chen, Environmental Protection Agency
James M. Landwehr, AT&T Bell Laboratories

4:00 p.m.
FLOOR DISCUSSION

4:45 p.m.
RECEPTION

THURSDAY, OCTOBER 2

9:00 a.m.
INTEGRATION OF EMPIRICAL RESEARCH:
THE ROLE OF PROBABILISTIC ASSESSMENT

Thomas Feagans, Decisions in a Complex Environment, Inc.

10:30 a.m.
BREAK

10:45 a.m.
DISCUSSANTS

Harvey Richmond, Environmental Protection Agency
Anthony D. Thrall, Electric Power Research Institute
Lee Merkhofer, Applied Decision Analysis, Inc.

11:30 a.m.
FLOOR DISCUSSION

12:00 p.m.
LUNCH

1:15 p.m.
STATISTICAL ANALYSIS OF POOLED DATA IN ECOLOGICAL AND
ENVIRONMENTAL WORK WITH SOME EXAMPLES

G. J. Babu, M. Boswell, K. Chatterjee, E. Linder,
G. P. Patil, & C. Taillie
Pennsylvania State University

2:30 p.m.
DISCUSSANT

Lloyd Lininger, State University of New York, Albany

3:00 p.m.
BREAK

3:15 p.m.
CONCLUDING PANEL DISCUSSION

Kinley Larntz, Washington State University

APPENDIX B: Conference Participants

ASA/EPA CONFERENCE ON STATISTICAL ISSUES
IN COMBINING ENVIRONMENTAL STUDIES
October 1–2, 1986

OMNI SHOREHAM HOTEL
WASHINGTON, DC

G.J. Babu
Pennsylvania State University
Department of Statistics
University Park, PA 16802

R. Clifton Bailey
U.S. EPA
6507 Divine Street
McLean, VA 22101

James C. Baker
U.S. EPA Region 8
999 18th Street, Suite 1300
Denver, CO 80202–2413

Ted O. Berner
Battelle Columbus Division
2030 M Street, N.W., Suite 700
Washington, DC 20036

M. Boswell
Pennsylvania State University
Department of Statistics
University Park, PA 16802

Robert N. Brown
Food and Drug Administration
200 C Street, S.W., MC–HFF–118
Washington, DC 20204

K. Chatterjee
Pennsylvania State University
Department of Statistics
University Park, PA 16802

Chanfu Chen
Lederle Laboratories
Building 60, Room 203
Pearl River, NY 01965

Chao Chen
U.S. EPA
401 M Street, S.W., RD–689
Washington, DC 20460

Jean Chesson
Battelle
2030 M Street, N.W.
Washington, DC 20036

Keewhan Choi
Exxon Corporation
Four Bloomingdale Drive, #517
Somerville, NJ 08876

Vincent James Cogliano
U.S. EPA
ORD, OHEA, CAG
401 M Street, S.W., MC–RD–689
Washington, DC 20460

Margaret Conomof
U.S. EPA
401 M Street, S.W.
Washington, DC 20460

Giles Crane
Department of Health
John Fitch Plaza, CN–360
Trenton, NJ 08625

John P. Creason
U.S. EPA
HERL/Biometry Division/MD–55
Research Triangle Park, NC 27711

J. Michael Davis
U.S. EPA
MD–52, ECAO
Research Triangle Park, NC 27711

Hari H. Dayal
Fox Chase Cancer Center
430b Rhawn Street
Philadelphia, PA 19111

Elizabeth A. Dutrow
U.S. EPA
401 M Street, S.W. (TS–798N)
Washington, DC 20460

David Eddy
Duke University
Center for Health Policy Analysis
Durham, NC 27706

Thomas B. Feagans
Decisions in a Complex Environment, Inc.
636 Wayland Place
State College, PA 16803

Bernice T. Fisher
U.S. EPA
1600 S. Eads Street, #525S
Arlington, VA 22202

Ruth E. Foster
U.S. EPA
401 M Street, N.W.
Washington, DC 20460

Michael E. Ginevan
9039 Sligo Creek Parkway, #1108
Silver Spring, MD 20901

John Goldsmith
U.S. EPA
Biometry Division, MD-55
Research Triangle Park, NC 27711

Noel P. Greis
Bell Communications Research
331 Newman Springs Road
Red Bank, NJ 07701

Gary F. Grindstaff
U.S. EPA
TS-798, 401 M Street, S.W.
Washington, DC 20460

Vic Hasselblad
Center for Health Policy Research
and Education
Duke University
Durham, NC 27706

Larry V. Hedges
University of Chicago
College of Education
Chicago, IL 60637

Robert W. Jernigan
U.S. EPA-SPB
American University
Washington, DC 20016

Woodruff B. Johnson
U.S. EPA
401 M Street, S.W., Room 223
Washington, DC 20460

Borko D. Jovanovic
University of Massachusetts
Department of Public Health
Amherst, MA 01003

Marvin A. Kastenbaum
The Tobacco Institute Inc.
1875 Eye Street, N.W.
Washington, DC 20006

Richard F. Kent
U.S. EPA
1 Scott Cirlce, N.W. #716
Washington, DC 20036

Kay T. Kimball
Oak Ridge National Laboratory
P.O. Box X, 4500S, MSF-260
Oak Ridge, TN 37831

Kathleen D. Knox
U.S. EPA
401 M Street, S.W., PM-223
Washington, DC 20460

Herbert Lacayo, Jr.
U.S. EPA
4520 King Street, #502
Alexandria, VA 22302

Emanuel Landau
American Public Health Assn.
1015 15th Street, N.W.
Washington, DC 20005

James M. Landwehr
AT&T Bell Laboratories
Statistical Models and Methods
Research Department
Murray Hill, NJ 07974

David Lane
University of Minnesota
270 Vincent Hall
Minneapolis, MN 55455

Kinley Larntz
Washington State University
Program in Statistics
Pullman, WA 99164-6212

Walter S. Liggett, Jr.
Center for Applied Mathematics
National Bureau of Standards
Gaithersburg, MD 20899

E. Linder
Pennsylvania State University
Department of Statistics
University Park, PA 16802

Lloyd Lininger
State University of New York-Albany
(U.S. EPA)
Albany, NY

Bertram D. Litt
U.S. EPA
OPP/Statistics (TS-769)
14502 Woodcrest Drive
Rockville, MD 20853

Rebecca A. Madison
U.S. EPA
401 M Street, S.W.
Washington, DC 20460

Sam Marcus
National Center for Health Statistics
13417 Keating Street
Rockville, MD 20853

Elizabeth H. Margosches
U.S. EPA (TS-798)
401 M Street, S.W.
Washington, DC 20460

Lee Merkhofer
Applied Decision Analysis
300 Sand Hill Road
Menlo Park, CA 94025

Barry I. Milcarek
Mobil Oil Corporation
150 East 42nd Street, Room 1324
New York, NY 10017

Patricia Murphy
U.S. EPA-Cincinnati
26 West Saint Clair
Cincinnati, OH 45268

Tom M. Murray
U.S. EPA
401 M Street, S.W.
Washington, DC 20460

C.J. Nelson
U.S. EPA (TS-798)
401 M Street, S.W.
Washington, DC 20460

Barry D. Nussbaum
U.S. EPA
EN-397F, 401 M Street, S.W.
Washington, DC 20460

G.P. Patil
Pennsylvania State University
Department of Statistics
University Park, PA 16802

Susan A. Perlin
U.S. EPA
401 M Street, S.W.
Washington, DC 20460

Lorenz R. Rhomberg
U.S. EPA (TS-798)
401 M Street, S.W.
Washington, DC 20460

Harvey Richmond
U.S. EPA
OAOPS, MC-MD12
Research Triangle Park, NC 27711

Wilson B. Riggan
U.S. EPA
HERL/Biometry Division/MD-55
Research Triangle Park, NC 27711

Frederick H. Rueter
CONSAD Research Corporation
121 North Highland Avenue
Pittsburgh, PA 15217

Joel Schwartz
U.S. EPA
1207 Fourth Street, S.W.
Washington, DC 20024

Judy A. Stober
U.S. EPA
26 West St. Clair
Cincinatti, Ohio 45268

Miron L. Straf
Committee on National Statistics
National Academy of Sciences/NRC
2101 Constitution Avenue, N.W.
Washington, DC 20418

David J. Svendsgaard
U.S. EPA
MD-55
Research Triangle Park, NC 27711

C. Tailie
Pennsylvania State University
Department of Statistics
University Park, PA 16802

Anthony D. Thrall
Electric Power Research Institute
P.O. Box 10412
Palo Alto, CA 94303

Harit Trivedi
Pennsylvania Department
of Environmental Resources
Bureau of Information Systems
Harrisburg, PA 17110

Alta Turner
Ebasco Services Inc.
160 Chubb Avenue
Lyndhurst, NJ 07071

Paul G. Wakim
American Petroleum Institute
1220 L Street, N.W.
Washington, DC 20005

John Warren
U.S. EPA
03023 (PM-223)
401 M Street, S.W.
Washington, DC 20460

Dorothy Wellington
U.S. EPA
401 M Street, S.W.
Washington, DC 20460

Herbert L. Wiser
U.S. EPA
Office of Air and Radiation
ANR-443, USEPA
Washington, DC 20460

Robert Wolpert
Duke University
Center for Health Policy Analysis
Durham, NC 27706

You-yen Yang
U.S. EPA
401 M Street, S.W.
Washington, DC 20460

as the latter, but in discussions it is sometimes applied to the former as well.[6] In this paper the usual definition of meta-analysis is both accepted and adhered to, with reanalysis of pooled data considered to be a third form of secondary analysis and not a form of meta-analysis. The distinction between meta-analyses and other analyses is then a clear-out distinction between analysis of empirical data in the case of the latter and analyses of results of empirical studies in the case of the former.

Returning to the rationale for doing meta-analysis, it has been cast along the lines that by using the deductive power of mathematics, in the form of statistical techniques, to integrate the results of a large set of studies it can be done more satisfactorily than by narration, just as the analysis of data in one of the original studies is done more satisfactorily by such techniques than by narration.[7] In the process of putting this rationale into some perspective, we can work toward specification of the function served by meta-analysis.

There are problems for both meta-analysis and secondary analysis of pooled data. For secondary analysis of pooled data, the raw data is not available in many cases. For meta-analysis, conventional statistical procedures are problematic for both statistical and conceptual reasons.[8] For both, study designs tend to differ in significant respects.

In the face of such problems, two-extremes are to be avoided. On the one hand, application of statistical methods is useful even when the conditions under which they are applied are not perfect in some sense. Also, methods more suitable for meta-analysis are being developed. The idea of developing and applying meta-analytic methods is unimpeachable, and not using them due to inertia or purism unwarranted.

On the other hand, it is important to avoid false dichotomies. Various means of informing policy decisions can be complementary rather than viewed as competitors Statistical methods are one powerful means of bringing the deductive power of mathematics to bear, means that serve an important function. But narration and other uses of mathematics which serve other functions need to be brought to bear as well.

The function in the decision-making process provided by meta-analyses is the application of statistical algorithms to the results of primary or secondary empirical studies. The purpose is to help deduce, infer, and consolidate implications of sets of studies. In so doing meta-analyses can reduce the amount of narration needed in state of information assessments.

The choice of algorithms to be applied in meta-analyses is subjective, but not arbitrary. Both substantive and statistical theoretical principals are applied where possible in making these choices. The choices are generally both judgmental and affected by substantive empirical content. Although two different persons might choose two different algorithms and/or sets of studies in a given case, any two persons would get the same result from correct application of the same algorithm to the same set of results.[9] Thus, although all three types of integration have both subjective and objective (intersubjective) aspects, meta-analysis has more objective aspects than the other two. For non-empirical Bayesian analyses some judgment would also enter as input to the algorithm in the form of subjective priors.

3.0 State of Information Assessments

Before making important policy decisions, it is useful to assess what is known and what is uncertain about the relationship between options and their consequences. Such assessments serve the ultimate purpose of the society maintaining as much control as possible over its future. Societal decison-making agents should make important decisions with what the society knows as a collective available to them. They should not make such decisions under the assumption that we know more about the connections between decison alternatives and their consequences than we do.[10]

Such assessments have been done in various forms and in all these forms, narration has played an important role. While meta-analyses can reduce the need for narration, they cannot eliminate this need. Non-formal exposition is essential for the task of interpreting the formal results of primary, secondary, and meta-analyses, particularly in interpreting their implications for the policy decisions at hand. Even for primary research, "often the statistical analysis is just a preliminary to the discussion of the underlying meaning of the data."[11]

State of information assessments (or scientific assessments) serve the function within the decison-making process of assessing the state of knowledge on which one or more important decisions are to be based. It is not the purpose of such assessments to add to the state of knowledge, either through empirical inquiry or through statistical analyses of data. Rather, the purpose is to assess the knowledge accumulated up to a point in time that is relevant to the decisions to be made at that time.

An important issue concernswhether there is a quantitative measure of the degree to which a given hypothesis, theory, or other proposition has been confirmed at a given time. Were such a measure to exist, it could play an important role in state of information assessments. The degree to which various theories about the shape and location of dose-response relationships were confirmed could be addressed, for example. Another example would be discussion of the degree to which the existence of a causal relationship between a given pollutant and a specified effect was disconfirmed by one or more negative studies.

This issue has received a great deal of attention from philosophers of science, measurement theorists, and statisticians interested in the foundations of their subject. Up until the mid-nineteenth century, attempts were made to construct theories of induction which guaranteed the truth of the conclusions to which their applications led. As soon as it became clear that some uncertainty or doubt about the truth of conclusions of inductive inferences is inevitable, methodologists began to consider scientific theories to be more or less probable, more or less worthy of rational belief.[12]Various attempts to reduce inductive logic to probability theory have followed. Despite efforts by such outstanding intellects as DeMorgan,[13] Jevons,[14] Pearce,[15] Keynes,[16] and Carnap,[17] all such attempts have failed.

These attempts have failed for a reason. Degree of confirmation or inductive support and probability are distinct concepts. The difference is subtle, but real and important. The failure to discern and explicate this distinction has bedeviled the history of these two (historically) conflated topics.

Most generally, the concept of probability has to do with the <u>balance</u> of favorable and unfavorable evidence; the concept of degree of confirmation has to do with the <u>amount</u> (and kind) of supporting evidence. If there is little evidence, favorable or unfavorable, for a proposition, probability assignments concerning the truth of the proposition can reasonably be near 0.5 . In contrast, by any reasonable account of confirmation, the degree of confirmation for the proposition is near 0.0.

Perhaps the most thorough attempt to develop a confirmation theory based on inductive logic was that of the philosopher of science, Rudolf Carnap. It is Carnap's terminology, "degree of confirmation," that is being used to describe the quantitative concept that is is important for the state of information integrative function. Carnap's work gave rise to thorough critiques of objectivist confirmation theories, and in his later years he began shifting toward a Bayesian point of view.[18]

The overoptimism concerning the possibility of a fully general and objective framework that pervades the historical attempts to develop inductive logic has been another obstacle to progress. Measurement of degrees of confirmation and probability are, most generally, more subjective than meta-analyses. The objectivity, in the sense of intersubjectivity, that is inherent to the meta-analysis function is unachievable for the other two integrative functions. We have used the appelation "meta-analysis" to name an algorithmic function since that term is gaining wide usage and the usage seems to roughly correspond to that function. (It should be kept in mind that from a decision-theoretic point of view, it is the functions, not the semantics, that are important.) One of the major keys to progress that has been made recently in confirmation theory is relaxation of the requirements of objectivity. [19,20] Although Shafer and Krantz have made significant progress in what we are calling confirmation theory, they do not recognize the distinction between probability and degree of confirmation. Hence, like so many before them, they refer to what we are calling degree of confirmation as "probability." Likewise, Cohen makes what appears to be a similar distinction in terms of "Pascalian probability" and "Baconian probability" [21]

4.0 <u>Probabilistic Assessment</u>

The role of probabilistic assessment in the decision-making process is to use whatever information and statistical analysis exists at the time the decision in question is to be made and relate the consequences the decision is designed to affect back to decision alternatives in a way that will achieve as much control as is feasible under the circumstances. Thus, the third integrative function, probabilistic assessment, uses the outputs of the first two integrative functions, meta-analysis and the state of information assessment. It is in turn an input to valuation analysis, decision analysis, and ultimately decision-making. Valuation analyses and decision analyses accomplish two other functions needed in support of decision making.[22,23]

Control is a meta-objective concerning the relationship between decision alternatives and primary objectives under uncertainty. Control in the sense meant here is analogous to the tightness/slack aspect of a steering mechanism and is not tied to any particular regulatory policy direction. The ultimate justification of the approach suggested to probabilistic assessment is that (in general) its greater generality gives more control.

Risk assessments are probabilistic assessments in which the consequences are adverse.[24] Risk assessments also include description of the seriousness of the adverse effects. The primary objectives in risk assessments are the avoidance of adverse health effects.

4.1 Probability

There are many ways the complex topic of probability can be addressed. In this discussion, we address the relationship between the levels of generality possible in making probability assignments and the perspective of the user of these assignments. This discussion will provide the basis for the selection of an approach to probabilistic (risk) assessment.

There are various possible interpretations of probability statements from the point of view of how they came to be made. These various interpretations have been much discussed for a long time. Three levels of generality fall out of all these discussions. At the lowest level of generality, probability assignments are made as the ratio of two nonnegative integers; if one of the integers is larger it is the denominator. This ratio may result from the application of a logical or relative frequency mathematical model.

The user of probability assignments only cares about how the probabilities were assigned in so far as it sheds light on how well a set of such assignments can be expected to predict in the probabilistic sense. Measurement of how well sets of probabilities predict involves two criteria, the criteria of calibration and resolution.[25] A canonical process of probability assignments can be defined in terms of these two criteria.[26] A canonical process of probability assignments is one in which the assignments are distributed randomly over the [0, 1] range (canonical resolution) and approach perfect calibration as a limit (canonical calibration). Many discussions of probability seem to implicitly assume a canonical process.

Canonicity does not necessarily obtain even at the lowest level of generality in the making of probability assignments. The phenomenon of ambiguity, illustrated by the Ellsburg Paradox experiments, demonstrates this fact.[27] The resolution of that paradox revolves around the theme of carefully distinguishing and analyzing the diverse perspectives of the maker and user of probability assignments.[28]

A second level of generality in making probability assignments is the degree of belief interpretation developed by the English philosopher, Frank Ramsey, [29] and the Italian statistician Bruno deFinetti.[30] This interpretation has been much used by decision analysts and Bayesian statisticians.[31] At this level of generality, probability assignments can be made by using a particular mathematical model, but only if the situation is deemed to justify the use of such a model. Many situations obviously do not. In such cases, final integration of the available information is done mentally and probability assignments are made judgmentally. Algorithmic devices, such as the mathematical/statistical models used in mathematical statistics in general and meta-analysis in particular, can be very useful aids in arriving at these judgmental assignments. Also, probabilistic models can be built which decompose the relationship in question so that the assignment can be derived from less difficult assignments. Thus, the amount of mental integration required is reduced to more manageable size.

At this level of generality, the fact that from the user's perspective canonicity may not obtain becomes critically important. It has been considered a positive characteristic of the Ramsey/deFinetti theory that despite its greater generality, and hence flexibility in application, its uninterpreted formal properties are equivalent to those of the narrower interpretations mentioned above--in other words, the usual probability mathematics applies. However, when judgmental assignments are made using states of information that will not give canonical probabilistic prediction, from a user's perspective, two probability assignments which are numerically equal should be interpreted differently.

In short, at the more general level, users of probabilities should interpret them in terms of both their numerical value and the state of information on which they are based. The divergency of probability from canonicity in general gives rise to the phenomenon of secondary risk.[32] The effect of secondary risk is to give less control to the user than he or she would have with canonical prediction. However, the user will have less control than he or she could have under the circumstances of the existing state of information if he or she misinterprets the probabilities to be canonical (in terms of probabilistic prediction) when they are not.

It happens that since both the numerical value of a probability and the state of information on which it is based are important, the Ramsey/deFinetti theory of probability is conceptually flawed, even from the perspective of applying it in making probability assignments. According to the theory, personal probabilities are revealed by sets of choices between pairs of bets. One bet in each pair is canonical by definition, and the other bet is not. But since states of information on which probabilities are based matter and since the states of information are different for the two bets within each pair, choices between each of these pairs of bets do not necessarily reveal personal probabilities. In general, the choices are affected by both the person's belief and the person's attitudes toward secondary risks.

Fortunately, there is an even more general theory of probability which does not have this problem. It is no accident that it is a more general theory that lacks this problem. Even though the Ramsey/deFinetti theory is more general than the first theory discussed above, it is not general enough. Both of the theories which are not general enough to serve as the basis of a normative framework for probabilistic (risk) assessment are available, as special cases of more general theory, for the domains to which they are general enough to apply. These domains are large sets of problems to which the methods of classical and Bayesian statistics, respectively, appropriately apply.

The more general theory of probability needed for the integrative function we are calling probabilistic assessment is the theory axiomitized by Bernard Koopman.[33,34] On this view, probability is an intuitive comparative relation that in general is only partially ordered.

The distinctive characteristic of an intuitive concept is that a large range of statements that employ a term signifying the concept can be understood and the meaning of the term cannot be explained in terms of more primitive concepts.[35] An intuitive concept can be applied correctly without using an explicit set of rules of application.[36] An intuitive concept is itself primitive.

Just as for the Ramsey/deFinetti theory, the key to eliciting probability judgments under the Koopman theory is to set up a comparison between the situation of interest and a canonical situation. The fact that under this theory probability is only a partially ordered comparative relation means that for some comparisons between the situation of interest and chosen canonical probabilities, the person making the probability judgments does not judge either to be more likely than the other. The fact that there is a range of canonical probabilities for which this is true in general means that in general, lower probabilities and upper probabilities are elicited rather than sharp probabilities.[37]

Axiomatically, the fact that in general probability assignments are not sharp means that an axiom which holds in the less general versions of probability (the additivity axiom) does not hold. The fact that in the Koopman theory the axioms associated with canonical situations do not apply is actually an advantage since there is less chance for confusion; that is, there is less chance that a user will interpret upper and lower probabilities to be canonical probabilities. The impression that it is an advantage of the deFinetti/Ramsey axioms that they are equivalent to the canonical (Kolmogorov) axioms is a misimpression generated by focusing on convenience of calculation for the producer rather than on the needs of the user. Assuredly, users have to learn how to interpret correctly the different kinds of output which result.

4.2 Normative Framework

The approach to probabilistic (risk) assessment described below was developed within the U.S. EPA Office of Air Quality Planning and standards risk analysis program by William F. Biller and the author.[38] Support analyses for national ambient air quality standards (NAAQS) must deal with enormous complexity. Thus, as general as possible an approach was needed and developed. Because of its generality, the approach serves well as a normative framework for probabilistic assessments. A normative framework should be as general as possible because it is relatively easy to reduce the generality of such a framework for the purposes of a specific application where the generality is either not needed or not advisable, but almost impossible to increase the generality of a framework within the context of a specific appliation.

The process of conducting a probabilistic (risk) assessment can be thought of as conducting a set of subprocesses that together make up the whole assessment: First, there is the subprocess of constructing a probabilistic risk model. Second; there is the subprocess of selecting those who are to make probability assignments, often substantive experts. Third, there is the subprocess of eliciting probability assignments. Finally, there is the subprocess of computing and presenting outputs.

4.2.1 Model Construction

The process of constructing a probabilistic risk model proceeds most rationally in accordance with a "back logic." The starting point is a set of adverse effects that regulatory policy could reduce. Starting with the consequences to be reduced, possible regulatory alternatives are identified which could reduce these consequences if adopted. Whether a causal relationship exists is uncertain in some cases. In all cases, the exact quantitative relationship between policy alternatives and consequences is uncertain. Both of these uncertainties can be handled formally within the framework.[40]

Assuming the relationship between policy alternatives and consequences of concern is decomposed for the purpose of reducing judgments to more manageable size, back logic is used to assure that the component models interface appropriately. For example, suppose the relationship between possible NAAQS's for a pollutant and adverse health effects to which the pollutant contributes is decomposed into three models:

1. standard-exposure model,
2. exposure-dose (physiological) model,
3. dose-response model.

It is the input-output structure of the dose-response model that is chosen first; then the exposure-dose model is chosen so that the dose given as an output is in the form of the input needed for the exposure-dose model. Similarly, the standard-exposure model is chosen so that the exposure output is the input needed by the exposure-dose model. The component models are chosen to give the most accurate output, with the form of the output a given constraint. This is the way the form of standards are best chosen, rather than relating standards directly to effects.

These component models are probabilistic models, so the overall model that relates decision options to the consequences to be affected is a probabilistic model. There is ordinarily no "correct" degree of fineness for the structure of the representation. Obviously a factor is the importance of the problem and the resources available for the assessment. A finer and therefore larger model costs more to build and implement. This is one of several questions of scale that must be decided before doing the assessment. The ideal would be to have alternative models of varying fineness of structure and then crosscheck and interrelate them. Such an approach would make the most use of indirect (background) information and coherence in the sense of consistency. Making maximum use of indirect information and coherence is most important for situations in which direct data is sparse.

Available data, statistical analyses of data, and indirect information are all considered in the process of constructing and choosing the probabilistic (risk) model. If two possible alternative models appear to be of equal merit as representations of the situation, but one has better information available to support probability judgments and other inputs, then it is preferred. If there is better data for the lesser representation, there is a tradeoff. If the better representation (model) is a refinement of the lesser representation, then the better representation is preferred since the data available for the grosser (lesser) representation constrains the inputs to the better representation.