

(

Title: A Data Fusion Framework For Meta-Evaluation of Intelligent Transportation System Effectiveness

Author(s): William M. Evanco

Dept.: JO90

Project No.: 0495 18B40A

Date: March 1996

Contract No.: DTFH61-95-C-00040

Issued at: Washington

Sponsor: Federal Highway Administration

ABSTRACT:

This study presents a framework for the meta-evaluation of Intelligent Transportation System effectiveness. The framework is based on data fusion approaches that adjust for data biases and violations of other standard statistical assumptions. Operational test characteristics that have a bearing on meta-evaluation methodology are identified in the context of the experimental paradigm. Data fusion approaches are presented for various types of measures of effectiveness and techniques for handling biases of various kinds are developed.

Keywords: meta-evaluation, operational tests, data fusion

TABLE OF CONTENTS

SECTION	PAGE
1 Introduction	1-1
1.1 Purpose of Study	1-2
1.2 Organization	1-2
2 Characteristics of Operational Tests	2-1
3 Meta-Evaluation	3-1
3.1 Requirements for a Meta-Evaluation Methodology	3-1
3.2 Experimentation to Assess ITS Services	3-1
4 Empirical Data and ITS Measurement	4-1
4.1 Data Types	4-2
4.1.1 Continuous Data Type	4-2
4.1.2 Integer Count Data Type	4-2
4.1.3 Categorical Data Type	4-2
4.1.4 Levels of Measurement	4-3
5 Statistical Biases	5-1
5.1 Biases to Internal Validity	5-1
5.1.1 Dilution	5-2
5.1.2 Contamination	5-2
5.1.3 Measurement Errors	5-2
5.1.4 Confounding Factors	5-3
5.1.5 Censoring	5-3
5.2 Biases To External Validity	5-3
5.2.1 Population Bias	5-3
5.2.2 Technology Bias	5-3
5.3 Comparability Bias	5-4
6 Statistical Inference	6-1
6.1 Continuous Data Type	6-1
6.2 Count Data Type	6-2
6.3 Categorical Data Type	6-3
6.3.1 Dichotomous Categorical Data Type	6-3
6.3.2 Polychotomous Categorical Data Type	6-4

SECTION	PAGE
6.4 Solution Methodologies	6-5
6.4.1 Maximum Likelihood Technique	6-5
6.4.2 Bayesian Analysis	6-6
7 Data Fusion	7-1
7.1 Data Pooling	7-1
7.2 Analyses of Controlled Experiments	7-2
7.2.1 Measures of Effect	7-3
7.2.2 Joint Likelihood Functions	7-4
7.2.3 Maximum Likelihood Estimates	7-4
7.2.4 Bayesian Analyses	7-4
7.3 Biases in Controlled Experiments	7-5
7.3.1 Dilution and Contamination	7-5
7.3.2 Measurement Errors	7-6
7.3.3 Confounding Factors	7-7
7.3.4 Censoring	7-8
7.4 External Biases	7-8
7.4.1 Population Bias	7-9
7.4.2 Technology Bias	7-9
7.5 Comparability Bias	7-10
List of References	RE- 1
Appendix A Experimentation in Other Fields	A-1
Appendix B Data Collection	B-1
Distribution List	DI-1

SECTION 1

INTRODUCTION

The Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991 calls for the deployment of surface transportation technologies for Intelligent Transportation Systems (ITS). A major purpose of ITS is to enhance the ability of this country to compete in the global economy. Among the objectives of the ITS program are the improvement of productivity and economic efficiency, the enhancement of transportation safety, the facilitation of traveler mobility, and the meeting of environmental concerns. In order to achieve these objectives, the ITS program is identifying advanced and emerging information, communications, control, and electronic technologies that have the potential to improve surface transportation.

The National ITS Program Plan calls for the development and ultimate deployment of twenty-nine interrelated user services. These user services are characterized in terms of the benefits for different users rather than in terms of their underlying technologies. The user services have been grouped into seven bundles as follows:

- Travel and transportation management
- Travel demand management
- Public transportation operations
- Commercial vehicle operations
- Electronic payment
- Emergency management
- Advanced vehicle control and safety systems

Among the initiatives included in the National ITS Program Plan is the conduct of operational tests related to the various user services. These tests are conducted for a prototype system on a scale smaller than full deployment over a relatively short period of time in a “real-world” (as opposed to a controlled laboratory or otherwise contrived) environment. One purpose of the tests is the evaluation of systems of ITS technologies that are wholly or in part beyond the R&D stage, but not yet ready for full deployment. Another purpose is the evaluation of user service benefits. Such evaluations will help to identify the more promising services and technologies (in terms of their impacts on ITS program objectives) for further development and deployment.

Multiple operational tests may be conducted for a particular user service. These tests are evaluated to provide one or more outcome measures of interest. When evaluation results for a number of operational tests become available, then it may be possible to synthesize results across tests to arrive at composite outcome measures. This synthesis process, called *meta-evaluation*, can provide information that is useful for determining the value of full scale deployment of an ITS user service. In addition, the utility of additional operational tests can

be determined by determining the statistical power of these tests and estimating the probability that they will deliver results in a specified range.

1.1 PURPOSE OF STUDY

The purpose of this study is to develop a data fusion framework for the meta-evaluation of ITS effectiveness. This framework provides a systematic capability for adjusting and synthesizing data from different tests to:

- Objectively synthesize evaluation measures and their uncertainties across multiple tests
- Identify the need for and characteristics of additional operational tests
- Estimate input parameters (and their uncertainties) for simulation models
- Provide ITS decision makers with information to guide decisions on the development and deployment of ITS systems

1.2 ORGANIZATION

In the next section, we discuss operational test characteristics that have particular bearing on meta-evaluation. In Section 3, desirable attributes of a meta-evaluation methodology are discussed and meta-evaluation is placed in the context of the experimental paradigm. Section 4 discusses the various types of empirical data and their measurement, while Section 5 discusses experimental statistical biases. Statistical inference as applied to meta-evaluation is discussed in Section 6. Finally, various approaches for data fusion are discussed in Section 7.

SECTION 2

CHARACTERISTICS OF OPERATIONAL TESTS

In this section, we discuss the differing characteristics of operational tests that may have an impact on the development of a meta-evaluation methodology. The methodology must be able to accommodate these differences.

Although different operational tests may involve the same ITS user service, the specific technologies providing the user service may differ among the tests. For example, in-vehicle route guidance systems (RGSs) may come in a variety of forms with varying accuracies for vehicle location and different means of providing dynamic route guidance.

Test settings are also expected to vary widely. Operational tests for a given ITS user service may be conducted by different teams, in a variety of geographic locations, and in different time frames. The mean characteristics of the populations under study (e.g., drivers using RGSs) may vary among the tests. Prototype ITS services being tested might differ from the one(s) that are ultimately deployed. Operational tests may also vary with respect to experimental design and be subject to different study-specific biases. Finally, operational tests may only provide indirect evidence of outcome measures of interest (e.g., close calls in place of accidents) and some of the tests may have gaps regarding certain outcome measures of interest.

An operational test may include one or more empirical studies. For example, an operational test for an RGS might involve an empirical study of yoked drivers (an equipped and an unequipped vehicle traveling the same origin/destination trip at the same time of day). The operational test may also support another empirical study of equipped and unequipped drivers not specifically paired with each other. Still a third empirical study may involve in-vehicle cameras to observe the details of driver interaction with the navigational device. All of these studies may provide data with regard to outcomes of interest such as travel time or safety. A relevant question then becomes how to merge results from these tests to draw meaningful conclusions about an outcome of interest.

Because of the relatively small scales of operational tests, system-wide impacts of large scale deployments may not be directly inferable from a test. For example, an operational test may not directly provide information about the impact of larger market penetrations of RGSs on traffic congestion or safety. Similarly, an operational test to evaluate weigh-in-motion (WIM) technologies that facilitate commercial vehicle inspections may only provide direct evidence of the time savings for an appropriately equipped commercial vehicle. The systemic effects of time savings of non-equipped vehicles (because of shorter queues at weigh stations) are difficult to measure directly. However, simulation or queuing models can be used for this purpose. In such cases, an operational test is used to determine values of some of the parameters for the calibration of a model. For example, operational tests for a

WIM technology can provide data regarding the service times for vehicles in a weigh station queue; this data can then be used in a queuing model to determine the systemic impacts of a specific level of WIM market penetration. When multiple operational tests or empirical studies have been conducted, then “best” estimates of these parameters can be obtained by merging results from the different sources. In addition, the “goodness” of the estimate can be inferred either from the parameter’s probability distribution or from its variance.

SECTION 3

META-EVALUATION

The traditional focus of statistics has been the analysis and interpretation of individual empirical studies. There are a variety of statistical methods appropriate for analyzing empirical data from different experimental designs. These methods allow the analyst to test hypotheses and to estimate parameters using approaches such as the maximum likelihood technique or Bayesian analysis. On the other hand, the task of adjusting and combining individual pieces of evidence from different studies has generally been left to subjective judgment. Typically, evidence from different empirical studies has been used to form an opinion-based impression regarding outcomes of interest.

The medical research community was among the first to recognize the need for a rigorous analytical methodology to facilitate the objective fusion of results from different empirical studies. Meta-evaluation techniques based on classical statistical methods have been developed for this purpose. A representative sampling of these techniques may be found in Wachter and Straf (1990). The techniques generally apply to empirical studies with a single outcome of interest each of which is conducted using a common experimental design. Moreover, it is generally assumed that there are no biases to either internal or external validity. These techniques involve either the pooling of data across empirical studies from which an outcome is computed or the combining of outcomes of a number of studies directly on the basis of different weights. Other approaches based on Bayesian analyses have been discussed by Eddy et al. (1990) and by Louis (1991).

3.1 REQUIREMENTS FOR A META-EVALUATION METHODOLOGY

Given the differences among operational tests cited in Section 2, there are a number of desirable characteristics that should be present in a meta-evaluation methodology. The methodology should be able to accommodate the incremental synthesis of evidence as it becomes available from operational tests. Another requirement for the methodology is that it be able to combine evidence from different tests not necessarily having a common experimental design. The ability to adjust individual pieces of evidence for biases is also a desirable feature. Finally, the methodology should be able to synthesize and incorporate indirect evidence for outcome measures, and to quantify and incorporate subjective judgments when necessary.

3.2 EXPERIMENTATION TO ASSESS ITS SERVICES

In order to further establish a basis for ITS meta-evaluation, it is necessary to understand the role of the experimental paradigm in the conduct of operational tests. Operational tests to

assess ITS services are intended to take place in a “real world” context. From purely the perspective of evaluation, ITS operational testing should ideally be conducted as an orderly and controlled experimental process. However, this ideal experimental paradigm is often difficult to achieve. Attempts to conduct fully controlled experiments may fall short of expectations. Since ITS systems involve human subjects, it is often difficult to create and maintain an ideal experimental environment. In Appendix A, we discuss the experimental paradigm as applied to the physical, biological/medical, and social sciences in order to provide a context for the following discussion of operational tests.

Population samples chosen for an operational test may include one or more experimental populations, which are subjected to various versions of an ITS technology, and a control population that does not use the technology. In some cases, a control population is not a part of the operational test. Instead, data may be collected with respect to a background population independent of the operational test. For example, an operational test may be conducted for a collision avoidance technology with only an experimental population. The impact of the technology may be discerned by comparing the accident rates of the experimental population with those of a background population collected from insurance statistics.

Experimentation as applied to ITS operational tests resembles the experimentation of the biomedical or social sciences rather than that of the physical sciences. The experimental environment may be imperfect and adjustments may be required to correct for biases and other problems. Quantitative evidence from either the operational test or external sources may be used to estimate these adjustments while, in some instances, the estimates may be subjective. In this latter case, sensitivity analyses may be conducted to determine the criticality of the subjective assumptions on outcomes.

A bias can occur within an operational test when there are differences between the experimental and control groups. For example, in an empirical study for an RGS the mean age of the experimental group using the navigational device might be different from the mean age of the control group. These age differences may partially account for differences in outcomes of interest so that the true effect of an ITS technology can be obscured. This is an example of a bias to **internal validity**.

Even if there are no differences between the control and experimental groups, their characteristics may be different from the population at large. For example, the drivers selected to partake in an RGS operational test may not be representative of the population as a whole. In this case, a comparison of outcomes between the experimental and control groups may not be a true representation of the effect of a technology in the general population, thus leading to a bias in **external validity**.

There may also be a bias to external validity when there are differences between a prototype ITS technology undergoing operational testing and the actual technology that will be

deployed. The impact of the ITS technology under operational test may be different from the deployed technology.

In each of the examples cited above, statistical corrections must be made, whenever possible, to adjust for problems in the experimental environment.

SECTION 4

EMPIRICAL DATA AND ITS MEASUREMENT

The purpose of an empirical study is to measure or otherwise characterize some empirical phenomenon in order to identify patterns, underlying laws of behavior, or explanations of outcomes. An important activity in the conduct of an empirical study involves the collection of data. Data collection and its processing is discussed in Appendix B. A datum is a measurement of an attribute of some subject of study, called a **unit of observation**. An attribute is a feature or property of interest associated with the unit of observation. The attribute may refer to either an outcome or an explanatory factor that is believed to affect an outcome for the unit of observation. The collection of attributes for a single unit of observation is called an **observation**.

For example, the unit of observation for an RGS operational test might be an individual driver on a specific origination/destination trip and the test may involve many such units of observation. The data collected for such a test can include outcomes such as travel times and travel distances, as well as explanatory variables such as the time of day the trip was undertaken, driver age, gender, and driving experience.¹

For the purposes of modeling, the outcomes and explanatory factors are represented by variables. Measurement is the process of assigning values to these variables through the empirical observation of each unit of observation. The variables may be related by functional relationships that can be expressed by means of equations. For example, the outcome variable of travel time may be assumed to be functionally related to the specific origination/destination trip, the time of day the trip was undertaken, driver age, gender, and driving experience. In this case, the outcome variable is said to be an **endogenous variable** while the explanatory variables are **exogenous variables**. If a particular functional relationship is hypothesized, then it can be empirically tested by using the data that was collected across the units of observation.

¹ The classification of a datum as an outcome is in part dependent on an implicit model of the relationships among the data and the purposes of the empirical study. In some cases, an outcome in one model will be the explanatory variable in another. Variables may be equivalently classified as exogenous (i.e., explanatory) or endogenous (i.e., explained). Whether or not a variable is endogenous then depends upon its use in a model or submodel.

4.1 DATA TYPES

Identifying a variable by its **data type** establishes the rules by which a variable is measured. In this study, we use a typology of three major data types, namely:

- Continuous
- Integer counts
- Categorical

As will be discussed below, the data type of an outcome or endogenous variable determines the types of statistical techniques that can be applied to its analysis.

4.1.1 Continuous Data Type

A variable that is continuous is one that is defined on all or part of the scale of real numbers. Travel time, expressed in minutes, or travel distance, expressed in miles and fractions thereof, are continuous variables defined between zero and infinity.

4.1.2 Integer Count Data Type

The count data type is appropriate for variables with an integer scale. Examples of such variables are the number of wrong turns or the number of “close calls” experienced by a driver during a specific origination/destination trip in an RGS experiment.

4.1.3 Categorical Data Type

The categorical data type is appropriate for variables whose values represent distinct categories. A variable may be dichotomous, having two categories, **or** polychotomous with more than two categories. An example of a dichotomous variable may be one measuring failure or success in some activity. For example, observing a group of drivers over some time period, a success might be defined for a specific driver as the absence of an accident during this period, while a failure would be the occurrence of one or more accidents in the same period. Polychotomous categorical variables are represented in the transportation literature by modal choices among multiple transportation alternatives Rassam et al. [1971]. For example, an automated traveler information system may affect the choices of air travelers who have four alternatives for getting to the airport: private automobile, taxi, bus, or rail.

The modal classification of transportation alternatives **is** an example of a categorical variable that has no implied underlying ordering to the categories. Each value is a distinct category that serves as a label for the category. A variable of this kind is called a **nominal** categorical variable.

An example of a categorical variable with an implied underlying ordering is one characterizing the severity of a vehicle accident. The categories might be: vehicle damage

only, personal injury, and fatality. In this case, it is possible to rank-order the categories according to some criterion, namely, accident severity. Each category possesses a unique position relative to the other categories. However, we do not know the “distance” between categories. A categorical variable with this property is an ***ordinal*** categorical variable.

4.1.4 Levels of Measurement

The different data types discussed above represent a hierarchy of levels of measurement. The categorical data type is the “lowest” level of measurement in the sense that the “higher” levels of measurement can be subsumed into the lower levels. Thus, a count can be subsumed in a categorical measure by grouping counts into two or more categories. For example, a population of drivers with an age distribution may be classified into three groups: 16-25 years, 26-40 years, and >40 years. Similarly, a continuous variable may be subsumed in a count or a categorical measure. For example, age is inherently a continuous variable if measured by years and fractions thereof. But for conceptual convenience, age is usually expressed as a positive integer.

SECTION 5

STATISTICAL BIASES

The meta-evaluation problem of combining results across operational tests is complicated by the expectation that operational tests may have statistical biases. In the ideal, an empirical study should be conducted such that its results, ***given the test circumstances, are*** a true reflection of the impact of a technology. However, the characteristics of the experimental and control groups may differ and these differences may, in part, contribute to the outcomes. Consequently, the true impact of the technology might not be discerned. In such cases, those factors that cause the evaluation measure to inaccurately reflect the impact of the technology in the test circumstances contribute to biases of ***internal validity***.

On the other hand, if the operational test circumstances differ from the expected deployment circumstances, then the operational test may lack ***external validity***. There may be no substantive differences between the experimental and control groups, but their characteristics may be different from the characteristics of the population that will ultimately be using the deployed ITS technology. In such cases, directly applying the results from the operational tests to the deployment population is not reasonable without first adjusting for biases to external validity.

When results from different operational tests are to be combined, it is first necessary to ensure that the individual operational tests are internally valid. Since the test circumstances for all of the tests may not be the same, combining the tests without further adjustments may lead to ***comparability biases***. In such cases, there are three courses of action. We may argue that the biases are small enough so as not to materially affect the evaluation measures. At the other extreme, we can argue that the biases are so large and uncorrectable as to make the operational test useless for application to the target circumstances. Or, we may try to adjust the experimental results to account for the biases.

For the last course of action, the factors potentially biasing an experiment must be identified. The directions of these biases need to be established, an attempt made to quantify their magnitudes, and a model built to incorporate the biases into the analysis. We discuss the various types of biases in greater detail in the following subsections.

5.1 BIASES TO INTERNAL VALIDITY

The difference between an outcome for an experimental group and that for a control group, called a ***measure of effect***, may be caused by biases to internal validity. Biases to internal validity include:

- Dilution of the experimental sample
- Contamination of the experimental sample
- Errors in measuring the outcomes
- Confounding factors
- Censoring (loss to follow-up)

To correct for biases to internal validity, models are required that relate the biased outcomes to the outcomes of interest (by way of additional parameters, if necessary). In the following subsections, we discuss examples of these various biases to internal validity.

5.1.1 Dilution

Dilution may occur when some of the individuals in the group using a technology do not actually use it. This situation can occur for any number of reasons. There may be willful intent on the parts of some individuals not to use the technology. In comparing accident mortality rates for drivers of seatbelt equipped vehicles with those without seatbelts, we must consider the possibility that some drivers may choose not to use their seatbelts. In this case, estimates of the reduction of vehicular accident mortality attributed to seatbelt use may be biased downward.

The equipment supporting a technology may be in “fail-mode” for certain periods of time. For example, communication failures or overloads may result in non-functional driver navigation aids during certain periods. In such cases, the driver may respond like a driver in a non-equipped vehicle.

5.1.2 Contamination

Contamination occurs when individuals who are not supposed to be exposed to a particular application of an ITS technology (e.g., in the control group) manage to get it on their own. Consider an experiment to determine the effects of different levels of training on the effective use of an ITS technology. One group may receive a high level of training, another group only a moderate level, and a third group no training at all. The groups with moderate or no training can introduce contamination by otherwise obtaining additional training on their own. In this case, the effect of training on outcomes would be biased downward.

5.1.3 Measurement Errors

The method used in the measurement of an outcome may not be accurate. If accident rates for a baseline group are obtained from insurance records, the rates may be biased downward since not all drivers may have insurance and not all accidents may be reported. Therefore, an accident rate comparison of an experimental group using a technology against a baseline group may lead to downward bias when measuring the effect of the technology on accident rates.

5.1.4 Confounding Factors

Confounding occurs when there are differences between the experimental and control groups other than in the application of an ITS technology. These differences may relate to the setting, to environmental differences, or to the characteristics of the individuals in the two groups. If the experimental group using the technology differs from the control group on the basis of some characteristics of the individuals and these characteristics can affect an outcome, then the effect of the characteristics might be falsely attributed to the effect (or lack of effect) of the technology. Selectivity bias is an instance of a confounding factor. For example, an individual may self-select or be selected by the experimenter with respect to one or more factors that can influence outcomes.

5.1.5 Censoring

Individuals in either an experimental or control group may drop out before completion of the experiment. If these individuals are different than the group as a whole with respect to their contributions to an outcome, then the outcome estimated from the remaining individuals may not represent that of the whole group. An experimental group for an RGS test may experience attrition as participants who experience difficulties using the equipment either discontinue its use or otherwise drop out of the testing program. This is an example of censoring and appropriate statistical techniques should be applied to account for it.

5.2 BIASES TO EXTERNAL VALIDITY

Differences that could affect outcomes between the operational test circumstances and those of the actual deployment may lead to biases to external validity. Biases to external validity also come into play when comparisons are made across experiments. These biases can be classified as population bias or technology bias.

5.2.1 Population Bias

Population bias occurs when there are differences between the experimental population of the operational test and the expected target population for the deployed ITS technology. For example, a population bias might arise if the group used to test a collision avoidance technology differs substantially from the general population of drivers in terms of age or gender composition. The measure of effect from such an experiment cannot be directly applied to the general population without adjustments.

5.2.2 Technology Bias

Differences between the ITS technology used in the operational test and the one that will ultimately be deployed may lead to a technology bias. If an operational test is conducted for an RGS based on GPS technology, then measures of effect, such as travel time, may be

different from those of a planned deployment of RGS using DGPS. Since the two systems differ with respect to locational accuracy, the probability for driver error (leading to wrong turns) may be greater for the GPS system, resulting in longer travel times.

5.3 COMPARABILITY BIAS

When conducting a meta-evaluation, the measures of effect across tests that are to be combined must be comparable. In order for these measures to be comparable, they must first be adjusted to correct for internal biases. Ideally, there should be no differences among the operational test circumstances that could affect the comparability of the measures of effect. If differences exist, then appropriate adjustments must be made to account for these differences, thus correcting for comparability bias. The measures of effect for the individual tests may then be adjusted for additional biases to external validity and combined, or they may be combined and then adjusted for these biases.

Comparability bias may be caused by differences in measurement across operational tests. These differences in many cases are definitional in nature. For example, several operational tests for RGSs may provide data on driver behavior such as the number of near accidents or close calls. These measures might be defined differently among the operational tests and, therefore, must be adjusted in order to achieve comparability.

Finally, differences across operational tests with respect to the time durations over which the tests are conducted can lead to a comparability bias if the definition of an outcome of interest depends on the time period. For example, the accident rates associated with a study conducted over a one year time period would differ from those of a six month study. The results of these operational tests again need to be standardized with respect to the time interval in order to make them comparable to each other. Moreover, the seasonal interval over which the six month study was conducted may require additional adjustments to the accident rates.

SECTION 6

STATISTICAL INFERENCE

In the absence of the various biases discussed previously, the merging of results across operational tests is relatively straightforward. Outcomes and measures of effect derived from operational tests are random variables. This randomness can be represented by means of parameterized probability distributions. Given empirical data on outcomes (and possibly explanatory variables) for units of observation from one or more operational tests, a likelihood function can be specified. The likelihood function represents the probability of observing the actual empirical outcomes of the units of observation associated with the operational tests. If the outcome for one unit of observation is statistically independent from the outcome of another (i.e., the usual case), then the likelihood function can be expressed as a product over the units of observation of the probability distributions evaluated at each outcome.

The likelihood function depends upon the parameters of the probability distributions. The values of these parameters determine the shape of the probabilities and hence of the likelihood function. Given the likelihood function, two approaches exist for the estimation of these parameters. The maximum likelihood approach selects parameters so as to maximize the likelihood function. This approach gives point estimates for the parameters and their associated standard deviations.

An alternative is to use the likelihood function as part of a Bayesian analysis to determine a distribution function for the parameters. The means and standard deviations of the parameters can then be computed from the distribution function.

If biases are present, then the likelihood function must be appropriately adjusted. In the next three subsections, we initially consider the case of no biases for each of the three data types. Likelihood functions are established for continuous, count, and categorical data types. Subsection 6.4 discusses two alternatives for the estimation of the parameters of the likelihood functions.

6.1 CONTINUOUS DATA TYPE

Outcomes that are continuous data types defined between $\pm\infty$ are often characterized by normal probability distributions. Many continuous outcome variables are distributed in such a way as to have at least an approximately normal distribution². For an ITS operational test, with an outcome z_i associated with the i th individual ($i=1, \dots, n$) in a sample of n individuals, the normal probability distribution is:

² Exceptions exist to this statement. For example, if an outcome variable were the mean time between events (such as failures of a system) then an exponential or Erlang distribution might be more appropriate.

$$\text{Pr}_N(z | \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{(z_i - \mu)^2}{2\sigma^2}\right) \quad (1)$$

where μ and σ^2 are the associated mean and variance of the distribution, respectively. Assuming that all of the observations are independently and identically sampled from the same normal probability distribution, then the likelihood function is the product of the probabilities for the individual observations (ignoring the $\sqrt{2\pi}$):

$$\begin{aligned} L(z | \mu, \sigma^2) &= \prod_{i=1}^n \text{Pr}_N(z | \mu, \sigma) \\ &= \frac{1}{\sigma^n} \exp\left(-\sum_{i=1}^n \frac{(z_i - \mu)^2}{2\sigma^2}\right) \end{aligned} \quad (2)$$

where $\mathbf{z} = (z_1, z_2, \dots, z_n)$.

To illustrate, an outcome variable might be the trip time associated with a sample of drivers using an RGS on a specified O/D trip. The trip time is a continuous variable, but it is defined on the interval between 0 and ∞ . If T_i is the trip time for the i th individual then the outcome z_i might be defined through the transformation $z_i = \ln(T_i)$. This transformation yields a z_i that is continuous and is defined on the interval $\pm\infty$, allowing the use of a normal probability distribution. In this case T_i is said to be log-normally distributed. In some instances, a normal distribution is a good approximation for a log-normally distributed outcome such as T_i . If the mean trip time is much greater than the standard deviation of the trip time, then trip time itself can be assumed to be normally distributed. In this case, the area under the left tail of the normal distribution for $T_i < 0$ would be extremely small.

6.2 COUNT DATA TYPE

If an outcome is expressed as a count data type, then it may be described by a Poisson probability distribution. For an ITS operational test, with an outcome z_i associated with the i th individual ($i=1, \dots, n$) in a sample of n individuals, the Poisson probability distribution is:

$$\text{Pr}_P(z | \lambda) = \exp(-\lambda) \frac{\lambda^{z_i}}{z_i!} \quad (3)$$

where λ is the expected number of counts. Assuming that all of the observations are independently and identically sampled from the same Poisson probability distribution, then the likelihood function is the product of the probabilities for the individual observations (ignoring the $z_i!$):

$$\begin{aligned}
L(z | \lambda) &= \prod_{i=1}^n \Pr_P(z | \lambda) \\
&= \exp(-n\lambda) \lambda^{\sum_{i=1}^n z_i}
\end{aligned} \tag{4}$$

An example of an outcome variable with the count data type is the number of near-accidents experienced by drivers using an RGS or the number of wrong turns that are made.

6.3 CATEGORICAL DATA TYPE

We have previously classified variables of the categorical data type as being dichotomous (i.e., two categories) vs. polychotomous (i.e., more than two categories), on one hand, and nominal (i.e., no implied ordering) vs. ordinal (i.e., implied ordering) on the other hand. In the next two subsections, we discuss the likelihood functions for the dichotomous and polychotomous categorical data types. The likelihood functions can be used for either nominal or ordinal categorical data.

6.3.1 Dichotomous Categorical Data Type

Suppose an ITS operational test has some categorical outcome that is dichotomous. The categories can be labeled by "A" and "B." The outcome for the i th individual ($i=1,2,\dots,n$) in a sample of n individuals can be represented by a binary variable, x_i , such that :

$$\begin{aligned}
x_i &= 0 && \text{if individual is in category A} \\
&= 1 && \text{if individual is in category B}
\end{aligned}$$

Let θ be the probability that an individual in the sample has $x_i = 0$. Then the probability for the individual to be in category A is

$$\Pr(x_i = 0 | \theta) = \theta$$

while the probability to be in category B is

$$\Pr(x_i = 1 | \theta) = 1 - \theta$$

Note that $\Pr(x_i = 0 | \theta) + \Pr(x_i = 1 | \theta) = 1$.

For the sample of n individuals, the probability that s of them have $x_i = 0$ (i.e., in category A) is given by a binomial distribution. Letting X be a random variable representing the number of individuals in category A, then:

$$\Pr_B(X=s | \theta, n) = \frac{n!}{s!(n-s)!} \theta^s (1 - \theta)^{n-s} \tag{5}$$

The likelihood function for the sample of n observations, s of which are in category A and $(n-s)$ of which are in category B, is then given by (ignoring the factorial terms):

$$L(X=s | \theta, n) = \theta^s (1 - \theta)^{n-s} \quad (6)$$

6.3.2 Polychotomous Categorical Data Type

If a categorical outcome is polychotomous with q categories, then the outcome for the i th individual ($i=1,2,\dots,n$) in a sample of n individuals can be represented by a multinomial variable, x_i , such that:

$$\begin{aligned} x_i &= 1 && \text{if individual is in the first category} \\ &= 2 && \text{if individual is in the second category} \\ &\vdots && \\ &= q && \text{if individual is in the } q\text{th category} \end{aligned}$$

Let θ_j be the probability that an individual i in the sample has $x_i = j$ where $j=1, 2,\dots,q$:

$$\Pr (x_i = j | \theta) = \theta_j$$

where $\theta = (\theta_1, \theta_2, \dots, \theta_q)$ and

$$\sum_{j=1}^q \Pr (x_i = j | \theta) = \sum_{j=1}^q \theta_j = 1 \quad (7)$$

For the sample of n individuals, the probability that s_1 are in the first category, s_2 in the second category, and s_q in the q th category is given by a multinomial distribution. Letting X_j ($j=1,2,\dots,q$) be random variables representing the numbers of individuals in categories j ($j=1,2,\dots,q$), then:

$$\Pr_M(X_1=s_1, X_2=s_2, \dots, X_q=s_q | \theta, n) = \frac{n!}{s_1!s_2!\dots s_q!} \theta_1^{s_1} \theta_2^{s_2} \dots \theta_q^{s_q} \quad (8)$$

The likelihood function for the sample of n observations is then given by (ignoring the factorial terms):

$$L(X=s | \theta, n) = \theta_1^{s_1} \theta_2^{s_2} \dots \theta_q^{s_q} \quad (9)$$

6.4 SOLUTION METHODOLOGIES

In the following subsections, we discuss alternative methods to estimate the parameters associated with the likelihood functions discussed above. There are two basic approaches for parameter estimation:

- Maximum likelihood technique
- Bayesian analysis

The approaches are discussed in terms of a generic likelihood function denoted by $L(\mathbf{X}|\Phi)$ where \mathbf{X} is a vector of observed test outcomes and $\Phi = (\phi_1, \phi_2, \dots, \phi_p)$ is the vector of associated likelihood function parameters.

6.4.1 Maximum Likelihood Technique

The maximum likelihood technique revolves around the selection of parameter values such that the likelihood function is maximized. The values of the parameters accomplishing this maximization are point estimates denoted by $\hat{\Phi}$. The parameters of the likelihood functions discussed in the previous sections are all continuous so that differential calculus techniques may be used to find their values.

It is generally convenient to work in terms of the natural log of the likelihood function defined by:

$$LL(\mathbf{X}|\Phi) = \ln[L(\mathbf{X}|\Phi)] \quad (10)$$

Since taking the logarithm of the likelihood function is a monotonic transformation, the parameter values that maximize the logarithm of the likelihood function will also maximize the likelihood function. The advantage of the logarithmic transformation of the likelihood function is that it converts products into sums which are easier to manipulate.

Thus, the parameter values that maximize the log-likelihood function are found by solving the equations:

$$\frac{\partial}{\partial \phi_i} LL(\mathbf{X}|\Phi) = 0 \quad \text{for } i = 1, \dots, p \quad (11)$$

for $\hat{\Phi}$. The solutions to equation (11) are guaranteed to maximize the likelihood function if the Hessian, which is the matrix of the second partial derivatives with respect to Φ of the log-likelihood function, is negative semidefinite. Equations (11) are generally not solvable through analytical means and must be solved by means of numerical iterative techniques. The Hessian may also be used to estimate the variances and covariances of the parameters as well as the associated confidence intervals. A discussion of these techniques is beyond the scope of this paper but the interested reader may refer to Kendall and Stuart [1961].

6.4.2 Bayesian Analysis

The Bayesian approach also starts with a likelihood function associated with the outcomes of some operational test. However, unlike the maximum likelihood technique, Bayesian analysis provides probability distributions for the parameters. The expected values and variances of the parameters can then be estimated from these distributions. The Bayesian approach requires that prior distributions be initially specified for the parameters. If the analyst initially has no information about the parameter values, then a non-informative prior distribution can be chosen to indicate his total ignorance about these values. On the basis of the likelihood function and the prior distribution, a posterior distribution can be computed that represents the probability distribution of the parameters. If results from a second test become available, the posterior distribution computed from the first test becomes the prior distribution that is input into a Bayesian analysis to incorporate the second test.

Let the prior distribution for the parameters be denoted by $\text{pr}(\Phi)$. If the parameters are independent, then

$$\text{pr}(\Phi) = \prod_{i=1}^p \text{pr}(\phi_i) \quad (12)$$

where $\text{pr}(\phi_i)$ is the prior distribution associated with ϕ_i , $i = 1, \dots, p$. Then the joint posterior distribution for the parameters is given by:

$$P(\Phi | \mathbf{X}) = cL(\mathbf{X} | \Phi)\text{pr}(\Phi) \quad (13)$$

where c is a normalizing constant chosen to ensure that the integration of $P(\Phi)$ over the parameter space equals unity.

Suppose the results from a second test subsequently became available. Then the posterior distribution from the first test (denoted by $P_1(\Phi)$) is used as a prior distribution for the second test. Let $L(\mathbf{X}_i | \Phi)$ be the likelihood function for the i th test ($i=1,2$). Then the posterior distribution of Φ for the second test is given by:

$$P_2(\Phi | \mathbf{X}_2) = c_2L(\mathbf{X}_2|\Phi)P_1(\Phi) \quad (14)$$

where c_2 is chosen to normalize $P_2(\Phi)$ to unity. Substituting equation (14) into (13), we then get:

$$P_2(\Phi | \mathbf{X}_1, \mathbf{X}_2) = c_2 c_1 L(\mathbf{X}_2|\Phi)L(\mathbf{X}_1|\Phi)\text{pr}(\Phi) \quad (15)$$

Equation (15) thus allows us to pool data across multiple experiments. The pooling of data is further discussed in Section 7.1.

SECTION 7

DATA FUSION

In the following sections, we discuss various approaches for the fusion of data across tests. Under ideal conditions, the following assumptions hold for two operational tests:

- **Assumption A:** The data of the two tests is identically sampled from the same probability distribution
- **Assumption B:** The observations of the two tests are independent
- **Assumption C:** There are no biases to internal validity, external validity, or comparability

In the next subsection, we discuss the pooling of data from two operational tests when all three of the above assumptions hold. In the subsequent subsections, we relax various combinations of the assumptions. The analysis of controlled experiments is discussed in subsection 7.2. Subsection 7.3 brings into play potential biases and the adjustments that can be made for them. In subsection 7.4, we discuss the fusion of data across empirical studies.

7.1 DATA POOLING

Suppose that we have two operational tests such that assumptions A, B, and C hold: Under these circumstances, we show below that for the likelihood functions discussed in subsections 6.1-6.4, the data from the individual operational tests can be pooled and treated as if they came from a single test.

For continuous outcomes, with outcomes denoted by $\mathbf{z}' = (z'_1, z'_2, \dots, z'_m)$ described by a normal distribution, the likelihood function for the two tests combined is just the product of the individual likelihoods given by:

$$\begin{aligned}
 L(\mathbf{z}, \mathbf{z}' \mid \mu, \sigma^2) &= L(\mathbf{z} \mid \mu, \sigma^2) L(\mathbf{z}' \mid \mu, \sigma^2) \\
 &= \frac{1}{\sigma^{n+m}} \exp\left(-\sum_{i=1}^n \frac{(z_i - \mu)^2}{2\sigma^2} - \sum_{i=1}^m \frac{(z'_i - \mu)^2}{2\sigma^2}\right) \quad (16)
 \end{aligned}$$

For outcomes that are integer counts following a Poisson distribution, the likelihood function for the two tests combined is given by:

$$L(\mathbf{z}, \mathbf{z}' \mid \lambda) = L(\mathbf{z} \mid \lambda) L(\mathbf{z}' \mid \lambda)$$

$$= \exp(-(n+m)\lambda) \lambda^{\sum_{i=1}^n z_i + \sum_{i=1}^m z'_i} \quad (17)$$

Once again, we see that the data from individual operational tests can be pooled and treated as if they came from a single test.

For dichotomous data, given a second operational test consisting of m units of observation, s' of which are in category A and $(m-s')$ of which are in category B, the joint likelihood function is given by

$$\begin{aligned} L(X=s, X'=s' \mid \theta, n+m) &= L(X=s \mid \theta, n) L(X'=s' \mid \theta, m) \\ &= \theta^{s+s'} (1 - \theta)^{n+m-s-s'} \end{aligned} \quad (18)$$

demonstrating that data from individual operational tests can be pooled and treated as if they came from a single test.

For polychotomous outcomes, suppose a second operational test has m units of observation, s'_1 of which are in the first category, s'_2 in the second category, and s'_q in the q th category. Then, the joint likelihood function is given by:

$$\begin{aligned} L(X=s, X'=s' \mid \theta, n+m) &= L(X=s \mid \theta, n) L(X'=s' \mid \theta, m) \\ &= \theta^{s_1+s'_1} \theta^{s_2+s'_2} \dots \theta^{s_q+s'_q} \end{aligned} \quad (19)$$

Thus, for the multinomial distributed likelihood function, data from individual operational tests can be pooled and treated as if they came from a single test.

7.2 ANALYSES OF CONTROLLED EXPERIMENTS

Suppose an operational test uses an experimental group and a control group for the purposes of comparison. These are called two-armed tests. For example, an RGS operational test may involve a group of individuals using the RGS technology compared to a control group not using the technology. In this case, we are effectively combining the results of two different experiments using groups from different populations (i.e., they differ by the application of the ITS technology). Thus, **Assumption A** discussed in the previous section no longer holds. However, **Assumption B** that the observations of the two tests are independent and **Assumption C** that there are no biases to internal and external validity (comparability is not applicable in this situation) are valid for a controlled experiment that is properly designed. We assume that we are dealing with such an experiment in the following. The analytical approaches discussed above can be easily adapted to accommodate such experiments.

7.2.1 Measures of Effect

For controlled experiments, one is less interested in outcomes than in comparisons between outcomes. Measures of effect (MOEs) are used to characterize these comparisons. Let ϕ represent a specific parameter from one of the likelihood functions discussed in subsections 6.1-6.3. Specifically, define:

$$\begin{aligned}\phi &= \mu && \text{if outcome is continuous} \\ &= \lambda && \text{if outcome is a count} \\ &= \theta && \text{if outcome is dichotomous}\end{aligned}\tag{20}$$

Then possible measures of effect (ε) are:

- Difference: $\varepsilon = \phi_e - \phi_c$ (21)

- Ratio: $\varepsilon = \frac{\phi_e}{\phi_c}$ (22)

- Percent difference: $\varepsilon = 100\left(\frac{\phi_e}{\phi_c} - 1\right)$ (23)

where ϕ_e is the parameter for the experimental group while ϕ_c is the parameter for the control group. For the special case of categorical data when $\phi = \theta$, we can define an additional measure of effect called the odds ratio given by:

- Odds ratio: $\varepsilon = \frac{\phi_e/(1 - \phi_e)}{\phi_c/(1 - \phi_c)}$ (24)

Since **Assumption C** holds, there are no biases to internal validity, so that differences between the experimental and control populations, for example, do not invalidate the measures of effect. A measure of effect, in this case, represents the true impact of an ITS technology on an outcome. Moreover, the test circumstances do not differ substantively from the deployment circumstances so that there are no biases to external validity. For example, the populations used in the experimental and control groups are representative of the population that will be using the deployed ITS technology. Thus, the measure of effect is an accurate representation of the effect expected if the ITS technology is deployed. Since at this point, no comparisons are being made across operational tests, biases to comparability are not relevant.

7.2.2 Joint Likelihood Functions

Since **Assumption B**, that the observations for the two populations are independent, holds, the joint likelihood function for the experiment is the product of the likelihood functions for each of the population groups:

$$L(\mathbf{X}_e, \mathbf{X}_c | \Phi_e, \Phi_c) = L(\mathbf{X}_e | \Phi_e)L(\mathbf{X}_c | \Phi_c) \quad (25)$$

7.2.3 Maximum Likelihood Estimates

Because of this independence, the estimation of the parameter vectors, Φ_e and Φ_c , can be made independently of each other. The estimates of specific parameters, ϕ_j ($j=e$ or c) can then be substituted into one of the MOEs defined above.

7.2.4 Bayesian Analyses

The posterior probability, given a likelihood function of the form (16), is

$$P(\Phi_e, \Phi_c | \mathbf{X}_e, \mathbf{X}_c) = L(\mathbf{X}_e | \Phi_e)L(\mathbf{X}_c | \Phi_c)pr(\Phi_e)pr(\Phi_c) \quad (26)$$

Suppose that we have some vector of measures of effect, ε , defined by

$$\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_s) = f(\Phi_e, \Phi_c) \quad (27)$$

where the number of measures of effect, s , is less than or equal to the number of parameters, p . The different definitions of measures of effect described in equations (12)-(15) have the property of involving only comparisons of a single parameter for the experimental and control groups. More complicated functions involving two or more distinct parameters for the experimental and control groups were not considered. Without loss of generality, we can let the s measures of effect be associated with the last s parameters in the vector Φ . Thus, the vector Φ_j ($j=e$ or c) can be written as:

$$\Phi_j = (\phi_{j1}, \dots, \phi_{j\ p-s}, \phi_{j\ p-s+1}, \dots, \phi_{jp}) \quad (28)$$

In this case, the individual measures of effect can be expressed as:

$$\varepsilon_i = f(\phi_{e\ p-s+i}, \phi_{c\ p-s+i}) \quad (29)$$

for $i=1, \dots, s$. Equations (20) can be solved for the $\phi_{e\ p-s+i}$ giving:

$$\phi_{e\ p-s+i} = g(\varepsilon_i, \phi_{c\ p-s+i}) \quad (30)$$

Substituting (21) into (19) for $j=e$ then gives:

$$\Phi_e = (\phi_{e1}, \dots, \phi_{e p-s}, g(\varepsilon_{p-s+1}, \phi_{c p-s+i}), \dots, g(\varepsilon_{cp}, \phi_{cp})) \quad (31)$$

Finally, substituting (22) into (17) the posterior probability can be expressed as a function of the vector $(\Phi_e', \Phi_c, \varepsilon)$ given by:

$$P(\Phi_e', \Phi_c, \varepsilon | X_e, X_c) = cL(X_e | \Phi_e', \varepsilon)L(X_c | \Phi_c)pr(\Phi_e', \varepsilon)pr(\Phi_c) \quad (32)$$

where:

$$\Phi_e' = (\phi_{e1}, \dots, \phi_{e p-s}) \quad (33)$$

If we are exclusively interested in the measures of effect ε , then we can consider the marginal probability for ε , which is obtained by integrating over the "nuisance" parameters (Φ_e', Φ_c) , and is given by:

$$P(\varepsilon | X_e, X_c) = \int_R P(\Phi_e', \Phi_c, \varepsilon | X_e, X_c) \quad (34)$$

where R denotes the region over which (Φ_e', Φ_c) is defined.

7.3 BIASES IN CONTROLLED EXPERIMENTS

Before any fusion of data can take place among empirical studies, it is necessary to adjust each of these studies for biases to internal validity. Part of **Assumption C** referring to internal biases is relaxed in this discussion. The appropriate adjustments for the different types of internal biases are outlined in the following subsections.

7.3.1 Dilution and Contamination

Dilution may occur when some of the individuals in the experimental group may not be fully exposed to the ITS technology, while contamination may occur if some of the individuals in the control group receive some exposure to the ITS technology. Let ϕ_i' represent a biased outcome parameter ($i=e$ for dilution and $i=c$ for contamination). Let α_i ($0 \leq \alpha_i \leq 1$) be a measure of exposure of group i to an alternative. Thus, if $i=e$ then α_e is a measure of the experimental group's exposure to some alternative other than the ITS technology. This alternative may be identical to that for the control group or it may be different from that of the control group. Similarly, if $i=c$ then α_c is interpreted as a measure of the control group's exposure to an alternative different from that of the control group. This alternative may be the ITS technology to which the experimental group is exposed or it may once again be

different. In either case, denote the parameter associated with this alternative as ϕ_0 . This parameter can be modeled as:

$$\phi_{i0} = \beta_i \phi_e + (1 - \beta_i) \phi_c \quad (35)$$

Thus for the experimental group ($i=e$), $\beta_e = 0$ means that the dilutants experienced the same exposure as the control group. Values of $\beta_e > 0$ imply that the dilutants experienced some level of technology between the experimental and control groups.

Similarly, for the control group ($i=c$), $\beta_c = 1$ means that the contaminants experience the same exposure as the experimental group and values of $\beta_c < 1$ imply that the contaminants experience some level of technology between the experimental and control groups.

The biased parameter (for $i=e$ or c) can then be written as:

$$\phi'_i = (1 - \alpha_i) \phi_i + \alpha_i \phi_{i0} \quad (36)$$

Equations (35) can be substituted into equations (36) which are then solved for ϕ_e and ϕ_c . The unbiased measure of effect is then given by $e = \phi_e - \phi_c$.

The parameters α_i and β_i can be estimated from available evidence or may be determined subjectively. If there is evidence on either of these parameters, then likelihood functions can be expressed as $L(x_\alpha | \alpha)$ or $L(x_\beta | \beta)$ and these likelihood functions can then be incorporated into a maximum likelihood or Bayesian analysis.

7.3.2 Measurement Errors

When measurement errors occur, the value assigned to an outcome may be different from the true value. If an accident rate is obtained for some baseline population, the measured rate may be an underestimate of the true rate because of under-reporting. If surveys, for example, provide independent information about the level of under-reporting, then this information can be used to relate the measured rate to the true rate by a multiplicative constant. More generally, measurement errors can be handled by means of a linear transformation of the measured outcome:

$$\phi' = \gamma + \omega \phi \quad (37)$$

where ϕ' is the measured outcome and ϕ is the true outcome.

For the special case when the outcome ϕ is a probability of an event occurring (e.g., an accident), then (37) can be written as:

$$\phi' = (\gamma + \omega) \phi + \gamma (1 - \phi) \quad (38)$$

The term $(\gamma + \omega)$ can be interpreted as the probability of an individual actually having an accident being reported as having had an accident, while γ is the probability of an individual not having had an accident being reported as one who did. These parameters may be estimated from empirical evidence or may be determined subjectively.

7.3.3 Confounding Factors

When the setting, environment, or the characteristics of the individuals in the experimental and control groups differ, adjustments must be made to "standardize" across the two arms of the empirical study.

A simple approach to handling confounding factors is to consider the hypothetical possibility that the individuals in the experimental group were instead placed in the control group. Let ϕ be the outcome associated with this group, while ϕ' is the outcome associated with the original control group. Then the relationship between the two outcomes may be taken to be:

$$\phi = v\phi' \quad (39)$$

where v is the factor that adjusts the control group for its differences from the experimental group. The parameter v may be estimated on the basis of available evidence or can be determined subjectively. Sensitivity analyses can be performed with various values of v .

A more comprehensive way to deal with confounding factors is to treat the differences between the two groups as covariates. For example, data may be available about the characteristics of the individuals participating in the empirical study. Denote these characteristics by vector $\mathbf{X} = (x_1, x_2, \dots, x_m)$. Then an outcome ϕ may be regarded as a function of these characteristics, $\phi = f(\mathbf{X})$.

More specifically, for continuous outcomes defined between $\pm\infty$, the parameter μ can be regarded as a linear function of the characteristics given by:

$$\mu = a_0 + a_1x_1 + a_2x_2 + \dots + a_mx_m \quad (40)$$

For outcomes that are counts, the parameter λ is a continuous variable defined between zero and $+\infty$. Taking the natural logarithm of λ converts it to a continuous variable defined between $\pm\infty$, so that an appropriate functional form for λ may be:

$$\ln(\lambda) = a_0 + a_1 \ln(x_1) + a_2 \ln(x_2) + \dots + a_m \ln(x_m) \quad (41)$$

Finally for dichotomous outcomes, a function of \mathbf{X} must be selected such that the parameter θ is limited between zero and unity. Such a function is the logistics function defined by:

$$\theta = \frac{1}{1 + \exp(a_0 + a_1 x_1 + a_2 x_2 + \dots + a_m x_m)} \quad (42)$$

This equation can be rearranged to yield an expression in terms of the logarithm of the odds ratio:

$$\ln\left(\frac{\theta}{1-\theta}\right) = a_0 + a_1 \ln(x_1) + a_2 \ln(x_2) + \dots + a_m \ln(x_m) \quad (43)$$

The functional forms in (40)-(42) can be substituted into the appropriate likelihood functions. The parameters a_0, a_1, \dots, a_m can then be estimated using either the maximum likelihood technique or Bayesian analysis.

7.3.4 Censoring

If the individuals who drop out of an empirical study have characteristics different from those who stay to completion, then observed outcomes for those who remain may be different from the outcomes if the group were kept intact. Let ϕ' be the outcome parameter for the group that stayed to completion and ϕ_d be the outcome associated with the fraction λ of the group that dropped out. Then the true outcome ϕ is given by:

$$\phi = (1 - \lambda)\phi' + \lambda\phi_d \quad (44)$$

The parameter ϕ_d can be estimated from available evidence or may be determined subjectively. More sophisticated approaches based on the characteristics of the individuals in the groups can be devised but the discussion of these approaches is beyond the scope of this study.

7.4 EXTERNAL BIASES

In the following subsections, the part of **Assumption C** referring to the absence of biases to external validity is relaxed. The population in an empirical study may not be representative of the population using the deployed ITS technology. In many cases, the ITS technology used in the operational test may not be identical to the ultimately deployed technology. In such cases, it is necessary to adjust outcomes and measures of effect to account for these differences.

7.4.1 Population Bias

If the empirical study population characteristics are different from those of the deployment population and information is available about the distributions of these characteristics in the populations, then adjustments can be made for these differences. The approach is similar to the one outlined in subsection 7.3.3 to handle confounding factors within an empirical study.

If this sort of detailed information is not available, then multiplicative models of the kind discussed in equation (39) can be used. The parameter v is the factor that adjusts some outcome ϕ' associated with an experimental population for its difference from outcome ϕ for the deployment population. This parameter either may be estimated on the basis of available evidence or may be determined subjectively. If empirical evidence is available, then a likelihood function for v may be established. For subjective estimates, sensitivity analyses can be performed using different values of v .

In the case of a two arm experiment, it may sometimes be better to work in terms of a measure of effect ϵ . Suppose we are evaluating a collision avoidance technology and are comparing an experimental group with a control group. Assume that there are no biases to internal validity or, if there are, they have been already resolved. Let θ'_i ($i=e$ or c) be the probability of having an accident for the i th group in the empirical study. On the other hand, θ_i ($i=e$ or c) is the probability of the i th group in the deployment population to have an accident. We note, however, that the population of the empirical study differs in age composition from that of the deployment population so that we would expect that $\theta'_i \neq \theta_i$ ($i=e$ or c). Consequently, a measure of effect that is the difference between the accident rates of the control and experimental groups (i.e., $\epsilon' = \theta'_c - \theta'_e$) would be expected to be different from the difference for the deployment population (i.e., $\epsilon = \theta_c - \theta_e$). However, if the rate of reduction of accidents is approximately independent of age, then a measure of effect that is the ratio of outcomes for the empirical study (i.e., $\epsilon' = \theta'_c / \theta'_e$) would be approximately equal to the ratio of the outcomes for the deployment group (i.e., $\epsilon = \theta_c / \theta_e$).

7.4.2 Technology Bias

If the ITS technology used in the empirical study is different from the technology to be deployed and measurements have been made regarding the impacts of these differences, then appropriate adjustments can be made. For example, an RGS system based on GPS guidance may result in more driver errors leading to wrong turns and hence more travel time than an RGS system based on DGPS. The number of wrong turns can be used as a variable to characterize the difference between the GPS and DGPS based systems. A model similar to equation (41) can then be specified to incorporate this variable as well as driver and other characteristics contributing to trip time. Sensitivity analyses can then be conducted with

respect to values of this variable to determine the potential impact of DGPS technology on trip time.

7.5 COMPARABILITY BIAS

Comparability biases need to be addressed when the results of different empirical studies are to be combined. Comparability bias is similar to external bias in the sense that the test populations, ITS technology, test settings, or environments must be comparable across empirical studies in order to combine their results. Consequently, the discussions of external biases also apply here. The populations across empirical studies may differ in terms of characteristics affecting outcomes and the technologies examined in the studies may not be identical.

In addition, when combining results of different empirical studies, differences in the definitions of outcomes may lead to a comparability bias. In such cases, it is necessary to adjust some outcome ϕ_1 for a test so that it is definitionally comparable to the outcome ϕ_2 of a second test. This may be accomplished by means of a functional transformation, $\phi_2 = f(\phi_1)$ a specific form of which is the linear transformation:

$$\phi_2 = a + b\phi_1 \quad (45)$$

where a and b are translation and scale parameters, respectively.

Another instance of comparability bias occurs when outcomes of empirical studies depend upon the time duration of the test. For example, accident rates from different empirical studies may need to be scaled in proportion to the time durations of the studies.

If a study involves the observation of accident rates for a variety of time periods, then a functional form for the accident rate in terms of the time duration can be expressed. One such functional form taken from equation (42) is:

$$\phi_i = \frac{1}{1 + \exp(a_0 + a_1 T_i)} \quad (46)$$

where T_i is the i th time duration. The parameter a_0 and a_1 can be estimated using a likelihood function. Once the parameters are estimated, then equation (46) could be used to estimate the accident rate for any time duration T .

LIST OF REFERENCES

1. Eddy, D. M., V. Hasselblad, and R. D. Shachter (1990), Bayesian Method for Synthesizing Evidence: The Confidence Profile Method, ***International Journal of Technology Assessment in Health Care* 6**, pp. 31-56.
2. Kendall, M. **G.** and A. Stuart (1961), The ***Advanced Theory of Statistics***, Volumes 1 and 2, Charles Griffm and Co.
3. Louis, T. A. (199 1), Using Empirical Bayes Methods in Biopharmaceutical Research, ***Statistics in Medicine* 10**, pp. 81 1-829.
4. Rassam, P., R. Ellis and J. Bennett (197 1), The n-Dimensional Logit Model: Development and Application, ***Highway Research Record* 369**, pp. 135-147.
5. Wachter, K. W. and M. L. Straf (1990), ***The Future of Meta-analysis***, Russell Sage Foundation, New York, New York.

APPENDIX A

EXPERIMENTATION IN OTHER FIELDS

A.1 Experimentation in the Physical Sciences

In the physical sciences, controlled laboratory experimentation plays a prominent role in extending knowledge about physical phenomena. An experiment is conducted by holding all but two variables fixed through careful regulation of the physical environment and then observing the dependence of one variable upon another. In this way, the functional dependency between the two variables can be identified. For example, the relationship between temperature and the pressure of a fixed volume of gas may be established by incrementally changing temperature over a specified range and observing the resulting pressure. Such an experiment is a controlled experiment since the volume of the gas is held fixed. Similarly, an experiment can be conducted to determine the dependence of gas pressure on the volume by holding the temperature fixed and varying the volume. The empirically observed functional dependency may be used in the formation of a theory of the phenomenon being observed. Conversely, the empirically determined dependency may be used to confirm an already existing theory.

A.2 Experimentation in the Bio-Medical and Social Sciences

This controlled experimentation paradigm has been carried over (with some modifications) to the biological, medical, and social sciences. In biology, the study of fruit flies, rats, or other living organisms replaces the inanimate physical systems studied in chemistry or physics. For example, a biologist may be interested in the impact of diet on the life expectancy of rats. In the ideal, he would like to conduct the experiment on a population of absolutely identical rats, splitting them into two groups—an experimental group containing rats on a “healthy” diet and a control group with rats on a “normal” diet. Unfortunately, such a population of absolutely identical rats cannot be found. Instead, the biologist may randomly split some population of rats into the two groups. If the initial rat population is large enough, any differences among the rats in the two groups will be statistically small. If a biologist either consciously or inadvertently selects the more healthy rats for one group and the less healthy ones for the other, then the results of the experiment will be biased. This is called a ***selectivity bias*** and when it occurs, it may have to be corrected by the use of sophisticated statistical methodologies.

In the social sciences and in medicine, the units of observation are often humans or groups of humans. Experimentation with humans presents many problems and the controlled experimentation paradigm is often difficult to apply in its rigorous form. Unlike rats, human experimental subjects cannot be caged. Hence, if humans are involved in a diet study, whether or not all the members of the “healthy” diet group keep to the diet is sometimes in question. Similarly, selectivity bias often plays a larger role in human experimentation. A

certain amount of voluntary participation is involved in many human experiments and whether or not an individual self-selects to partake in an experiment may be driven by factors not under the experimenter's control.

An experiment involving a control and an experimental group is called a ***two-armed experiment***. In some cases, the control group may be eliminated. If an outcome of interest has already been measured for the general population, then the general population may, in essence, be regarded as a "control group." The advantage of such an approach is that the limited resources for an experiment can be focused on collecting data from a larger experimental group. The outcome in the experimental group can then be compared to that of the general population. Of course, to avoid a biased comparison, the experimental group should have a composition similar to that of the general population.

A.3 Naturally Occurring Experiments

Attempts to conduct a controlled experiment are sometimes not fully successful resulting in only a partially controlled experiment. Moreover, in some contexts, it is either impossible or otherwise impractical to apply the controlled experimental paradigm. Astronomy is a prime example whereby researchers must rely on physical phenomena that occurred far distant in space and time. The possibility of controlled experimentation in such a situation is scant. Another example is in economic research. The feasibility of experimentation on large scale economic structures or even on smaller collections of individuals engaging in economic activity is limited. In these cases, researchers collect data on "naturally occurring experiments" and exploit the variation in the data for their statistical analyses. Multivariate statistical methodologies allow the researcher to "control" for confounding variables and to identify the dependencies among the variables of interest. This approach is opportunistic in the sense that data variation occurring naturally for some physical, biological, or social phenomenon is exploited through the use of these statistical methodologies.

APPENDIX B
DATA COLLECTION

Raw data is the immediate result of the data collection process. This raw data must be processed to yield refined data. The refined data may then be further transformed, combined, and aggregated to provide "views" of progressively coarser granularity. This data collection and refinement process is depicted graphically in Figure 1.

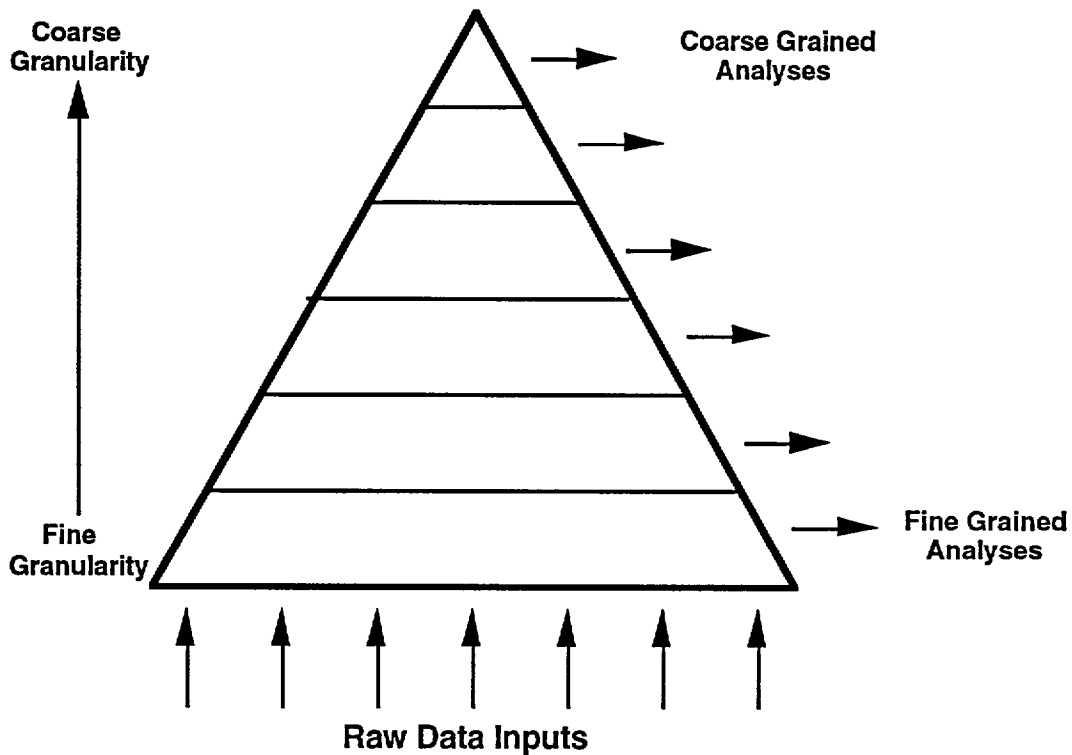


Figure 1: Data Refinement Pyramid

Raw data enters the bottom of the pyramid in the form of data collection forms or inputs from automated data capture mechanisms. The raw data from the various sources must be validated, further processed, and fused before it enters into the lowest level of the pyramid as refined data. This first level of refined data is then input to the second level for additional processing, validation, fusion, and aggregation. The second level thus consists of data at a coarser level of granularity than the first level. This process is continued until the top of the

pyramid is reached, at which point the data is at the coarsest level of granularity. Analyses may be performed on any of the levels of granularity as shown in Figure 1.

As an example, an operational test may be conducted to collect data for individual drivers partaking in a test of an RGS. At the first level, detailed data may be maintained for link times and distances, driver behaviors such as eye movements and their times of occurrence, detailed information about “close calls” and driver errors, and their associated times, the real-time inputs of data from a traffic management system, and various characteristics of the driver.

The second level may involve data for each driver which is aggregated over link times to provide origination/destination times and over link distances to provide origination/destination distances. Moreover, this level may include only summary statistics about driver behaviors and data inputs from the traffic management system. This process of data reduction and aggregation may continue to the top level of the pyramid at which point the data may include aggregate statistics such as the number of drivers in the experimental and control groups, the origination/destination trip traveled, their average travel times and travel distances, the average number of “close calls,” etc.

The data from the bottommost level of the pyramid supports, for example, micro-analyses of man-machine interface issues for which a second-by-second accounting of driver behavior is relevant. At an intermediate level of aggregation, the data could support detailed analyses of the determinants of travel time or distance, while at the topmost level various “comparison” questions can be addressed regarding the differential impacts of the RGS system versus its not being used.

The data pyramid does not necessarily imply that the data at any given level is maintained in a physical database. A given level may be produced on an “as-needed” basis by applying to the lower levels software utilities that appropriately aggregate and process the data. This processing may include the generation of composite or derived measures. A specific analysis may even require data from one of the higher levels to provide contextual information. For example, aggregated traffic management information can be used for a driver level analysis to provide a measure of traffic congestion during a test.

We expect each empirical study to have a data pyramid of the sort discussed in Figure 1. For empirical studies concerned with a single ITS user service, the raw data collected may differ in substantive ways. Differences may occur in the sizes as well as the characteristics of the populations in the study, the settings in which the study takes place, the experimental designs used to collect the data, and the specifics of the ITS technology being tested. We must be able to account and adjust for these differences if a reasonable basis for meta-evaluation is to be found.

DISTRIBUTION LIST

INTERNAL

JO90

K. J. Biesecker
R. Bolczak
J. R. Cerva
M. D. Cheslow
D. L. Dion
K P. Dopart
W. M. Evanco (10)
R. A. Glassco
S. G. Hatcher
C. W. Kain
K M. Lamm
R. K. Lay
M. F. McGurrin
A. A. Mertig
J. L. Milner
G. G. Nelson
R. G. Nystrom
V. M. Patel
J. R. Peterson
A. T. Proper
D. L. Roberts
A. E. Salwin
A. M. Schoka
D. E. Shank
K. E. Wunderlich
ITS File (15)
JO90 File (2)

JO10


P. D. Bergstrom

Transportation Data File (2)
Records Resources (3)

EXTERNAL

Mr. Michael Frietas, HVH- 1
Michael Halladay, HVH- 1
Dr. Christine Johnson, HVH-1
Mr. Jeff Lindley, HVH-1
Dr. Joseph Peters, HVH-1
Federal Highway Administration
400 7th Street, S.W.
Washington, DC 20590

Approved for Project Distribution:



M. F. McGurrin
Program Manager, 0495 18B40A