#3

Accurate, detailed, representative data on U.S. automotive accidents can be obtained only through a carefully designed and controlled data collection system. Police-reported data lack consistency and sufficient detail, and thus are of limited value for analytic purposes. On the other hand, multidisciplinary studies of selected accidents have little generalizability. The data required for effective cost-benefit studies of nationwide phenomena must be obtained by means of a carefully designed sampling plan. Staff members of the University of Michigan Highway Safety Research Institute, in consultation with the Sampling Section of the University of Michigan Institute for Social Research, have designed the sampling plan discussed in this report. The plan provides a means of scientifically and economically obtaining the needed data.

Leslie Kish
Professor
Research Scientist,
ISR

UM-HSRI-SA-75-14 (Volume 1 of 2)

# DESIGN FOR
# NASS: A NATIONAL ACCIDENT SAMPLING SYSTEM

J. O'Day
A. Wolfe
R. Kaplan

July, 1975

Prepared For

National Highway Traffic Safety
Administration
U.S. Department of Transportation
Washington, D.C. 20590
Under Contract No. DOT-HS-4-00890

Prepared By

Highway Safety Research Institute
The University of Michigan
Ann Arbor, Michigan 48105

**Technical Report Documentation Page**

| | | |
|---|---|---|
| 1. Report No. | 2. Government Accession No. | 3. Recipient's Catalog No. |

4. Title and Subtitle

DESIGN FOR NASS: A NATIONAL ACCIDENT SAMPLING SYSTEM

5. Report Date

July, 1975

6. Performing Organization Code

7. Author(s)

J. O'Day, A. Wolfe, R. Kaplan

8. Performing Organization Report No.

UM-HSRI-SA-75-14<br>(Vol. 1 of 2)

9. Performing Organization Name and Address

Highway Safety Research Institute
The University of Michigan
Ann Arbor, Michigan 48105

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

DOT-HS-4-00890

13. Type of Report and Period Covered

Phase I
June, 1974-July,1975

12. Sponsoring Agency Name and Address

National Highway Traffic Safety Admin.
U.S. Department of Transportation
Washington, D.C. 20590

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

A design is presented for a national accident investigation program based on sampling theory. By limiting the number of investigations within a strict sampling plan it is possible to record sufficient detail about each accident to produce national estimates of injury, property damage, and other accident characteristics which will be useful in cost-benefit analyses. The system described has three major facets--a program for continuous acquisition of data of a random sample of all towaway-pedestrian-bicycle-motorcycle accidents occurring in the U.S., a program for occasional acquisition of additional data on selected topics quickly and on-call, and a program for conducting in-depth or multidisciplinary accident investigations for accidents of particular interest.

While alternative approaches are discussed, the system recommended consists of 35 primary sampling units distributed throughout the 48 contiguous states. The design is complete and the system is ready for pilot implementation. Full implementation is possible over a period of three years.

17. Key Words

Data System Design
Accident Data
Sampling Plan
NASS, CSS, QRS, MDAI

18. Distribution Statement

Unlimited

| | | | |
|---|---|---|---|
| 19. Security Classif. (of this report) | 20. Security Classif. (of this page) | 21. No. of Pages | 22. Price |
| Unclassified | Unclassified | 359 | |

Reproduction of completed page authorized

★ U. S. GOVERNMENT PRINTING OFFICE : 1973 725-504/328

## CONTENTS

APPENDICES (in Volume 2)

# LIST OF TABLES

## LIST OF FIGURES

Figure Number

# SUMMARY

This report presents the design of a National Accident Sampling System (NASS). The purpose of this system is to collect highway accident data nationwide in accordance with a statistical plan that ensures the reliability, validity, and representativeness of the data as a precise, continuing microcosm or model of highway accidents occurring in the U.S. This NASS design has been developed by the Highway Safety Research Institute and the Institute for Social Research, The University of Michigan, under contract to the NHTSA. The system as now developed is ready for testing by means of a pilot program.

The NASS has been designed to house three sub-systems, each with a unique purpose. The first of these is a continuous sampling sub-system (CSS) which provides for a continuous tracking of a national sample of serious accidents with a substantial amount of detail. The second is a quick reaction sub-system (QRS) which permits the director of the system to quickly specify and carry out a data collection program for a specific purpose. The third is a full multidisciplinary accident investigation capability (MDAI), similar to that currently in use.

A system such as NASS is needed to overcome problems inherent in present accident data collection programs. Police-reported data provide quantity without adequate detail, precision, or consistency. MDAI-team data provide quality without adequate representativeness for making valid and reliable inferences concerning accident phenomena nationally. The NASS design described here solves those problems by providing for data collection by 35 teams in 35

precisely defined Primary Sampling Unit areas in the U.S., with the teams established, trained, directed, and monitored by a central headquarters unit. As now designed, the NASS system complements the current NHTSA Fatal Accident Reporting System (FARS) and the currently developing National Accident Report System (NARS) by providing significantly more detail concerning crash damage and injuries than could be obtained from the planned larger NARS and FARS data sets.

To obtain 15,000 to 20,000 representative accident cases annually (all that are needed for a valid representation of most national phenomena), NASS employs a sample design briefly describable as a multi-stage, controlled-probability, cluster design. This controls the selection of accident cases on the basis of population, gasoline sales, and geographic regions. The 35 Primary Sampling Units include Los Angeles, Chicago, and New York, with the other 32 PSUs equally divided in four regions of the U.S. (Northeast, Midwest, South and West). Three population density strata are represented: urban (central counties of Standard Metropolitan Statistical Areas); suburban (non-central counties within SMSAs); and rural (non-SMSA counties).

The population of accidents to be sampled by NASS includes all fatal accidents, all accidents in which at least one vehicle is towed from the scene, and all police-reported pedestrian, bicycle, and motorcycle accidents. Accidents selected for inclusion in the sample will be reported on in detail by trained investigators who will visit the site, inspect and measure the vehicle, and when necessary, interview the driver, other occupants, attending physicians, etc. The frequency and severity of injuries occurring in non-towaway and non-reported accidents would be estimated through separate general public surveys coordinated with the NASS accident investigation program.

The NASS design integrates the strengths of human investigators and digital computers to assure collection of incontrovertible data essential for effective evaluation of present countermeasure programs and the development of new programs. The computer programs necessary for data processing, inputs, and analyses are complete and ready. The forms for field data collection are ready. The organization plan for establishing PSU points, field teams, and functions of the units is complete and is herein described. Thus it is recommended that a pilot program consisting of four or five PSU teams be initiated during the next year. Current MDAI teams would be used to implement this pilot program. That pilot program would then be expanded in the following year to include 16-18 PSU teams that would begin to provide nationally useful information. It is envisioned that the full recommended system of 35 PSU teams could thus be made operational within a three-year period.

# 1. INTRODUCTION

This report describes a national traffic accident sampling system designed to obtain accurate, detailed, and adequately precise data on highway accidents in the United States. The data obtained by means of this system can be used to identify national problems which need attention, to determine the effects of some highway safety standards, and to provide a quantitative basis for designing and evaluating countermeasure programs. The system consists of (1) a continuous sampling sub-system for investigating approximately 15,000 accidents per year in detail; (2) a quick-reaction sub-system operating in the same frame and capable of compiling special data on a chosen subject quickly with the same sampling precision as in the continuous system; and (3) a multidisciplinary accident investigation subsystem for full, in-depth investigations of selected accidents.

The National Highway Traffic Safety Administration has supported several accident investigation programs since its formation in 1966, all concerned with developing an understanding of the problems of highway safety. A major part of the NHTSA effort has been in support of multidisciplinary accident investigation teams which have operated in many parts of the U.S. and which have addressed several specific study topics. In 1972, The University of Michigan's Highway Safety Research Institute undertook for NHTSA a study* of the value of the MDAI data for drawing statistical inferences about highway crashes in the U.S. One of the outputs of that study was a recommendation that the in-depth

*Statistical Inference from Multidisciplinary Accident Investigation, DOT-801 11, August, 1974, J. O'Day, *et al.*

accident investigation programs be modified toward a national sampling system to permit stronger inferences to the national population. Another recommendation was that detailed data collection be continued, so that both types of information--representative and in-depth--would be available to designers, evaluators, policy-makers, and planners.

The purpose of this report is to describe the developed design, to detail some design options at this point, and to lay out a program of analysis and pilot testing that will reduce operational uncertainties and thereby facilitate implementation of a full system of data collection and analysis.

The design presented here employs what has been learned in nearly nine years of Department of Transportation experience in acquiring knowledge about the national accident population. The design employs well-developed methodologies for investigation and for precise reporting, and it applies modern computer technology to provide the system responsiveness necessary for assuring collection of accurate and representative data. While the basic investigations remain the responsibility of human investigators, the system is designed to link the unique strengths of humans and computers to obtain accurate, precise, detailed, representative data.

The technical problems addressed in this report fall into two general categories: (1) sample design, such that the collection system can produce data which are representative of the U.S. accident population in a definable way; and (2) organizational and operational design, such that the data (to be acquired under a defined sampling program) are complete, precise, and accurate enough to permit valid conclusions to be drawn. This dichotomy of sample (or mathematical) design of the system and the operational (and organizational) design will be addressed in detail in many parts of this report.

The design of a complex system is a continuing, semi-reflexive process. It can be viewed as a sequence of (1) an idea, (2) a conceptual design, (3) a mathematical design, (4) an operational-functional design; and (5) a pilot test, etc.--with each step contributing to other steps in the process (as shown in Figure 1). This report describes the major features of the NASS design. It introduces and describes a conceptual design based on the elements shown in the "background" and "constraints" boxes. This conceptual design defines a population of accidents of interest, defines a sampling frame, describes methods of handling the data, and presents some pre-pilot experience (from an ongoing restraint system evaluation program). The system-design choices that remain open at this point can best be made with information obtained from pilot operations. The system in operation will be productive; the choices which remain are not critical to the basic design. Therefore the appropriate next step is a formal pilot program to test the NASS design.

The remainder of this report presents the background for the present study (Section 2) and an overview of the system (Section 3). The sampling design at the primary (national), secondary, and tertiary levels is discussed in Section 4. The steps to the selection of the actual sample are detailed in Section 5. Data collection, handling, analysis, and reporting are described in Sections 6-8. The command and control functions within the system are discussed in Section 9, and monitoring (which is directly related to command and control) is discussed in Section 10. Section 11 discusses the functional organization. Section 12 suggests a schedule for implementation of the

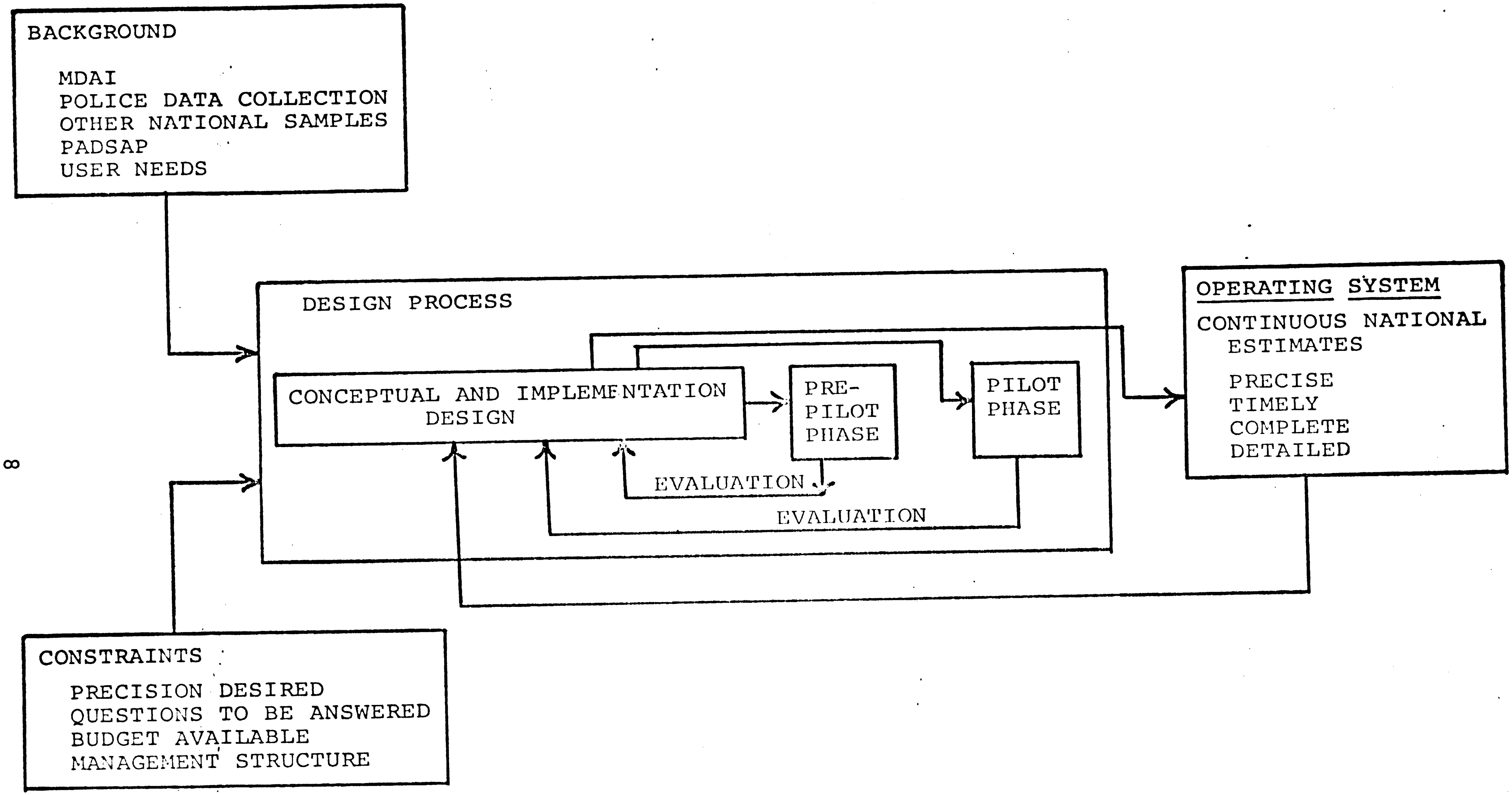


FIGURE 1.1. THE PHASES OF DEVELOPMENT OF THE NASS

system. And Section 13 presents the costs for the operating system through the various phases.

Appendices present more detailed information on various aspects of the system.

# 2. BACKGROUND

Highway safety policymakers, standards writers, evaluators, and designers seldom can use and do not want data in its raw forms. But they need the results of analyses of high-quality raw data. The combination of valid data and consistent analysis is particularly valuable as an input to cost-benefit computation--in the development of standards, of alternate designs for vehicles, and in the adoption of countermeasures by police or other governmental agencies.

Administrators and planners need accurate and reliable answers to such questions as:

-Are air bags more effective than shoulder harnesses in reducing injury?

-Is the relationship between accident causation and variability in vehicle-handling-and-braking characteristics strong enough to justify the imposition of performance standards in this area?

-Is driver education effective in reducing accident involvement?

-Can particular groups of drivers be identified as over-involved in accidents, and thus be targets for countermeasure development and application?

-What is the rate of restraint system usage in cars of several (recent) model years at the time they are involved in collisions?

-Are large trucks "overinvolved" in injury-producing accidents?

-What kinds of injuries are sustained in side-impact collisions? What is the source (i.e., the contact points) for these injuries?

-How many in-car injuries are incurred nationally at each code level of the Abbreviated Injury Scale?

-What is the frequency of roadside obstacle crashes, and what are the distributions of injury severity and type of obstacle?

While answers to some of these questions lie in simple tabulations of data (e.g., the injury distribution on the Abbreviated Injury Scale for car occupants), others require considerable reasoning and multivariate analysis. For example, the effectiveness of one type of restraint compared with another may well be modified by differences in the ages of drivers or the sizes or weights of the cars used.* An effective accident data system must provide not only valid data but also the analytical capabilities for using that data effectively.

## 2.1 Existing Data and Programs

Accident data in one form or another have been available for many years. The National Safety Council has compiled and published state and city fatal traffic accident statistics, and has estimated national accident totals annually in its "Accident Facts" booklet. Following enactment of the 1966 Transportation Act, most states have shifted toward uniform (within the state) accident reports and to digital processing of the acquired accident data.

Within the NHTSA, police-reported data were first compiled in a digital file identified as the National Accident Summary for the year 1968, and subsequent data sets of the NAS have provided a relatively complete but undetailed set of police-reported accidents. NHTSA's present Fatal Accident Reporting System, and the planned extension to a National Accident Reporting System, provide for more consistent reporting

*See, for example, "How Much Safer Are You in a Large Car", J. O'Day and R. Kaplan, SAE Paper 750116.

of selected accidents from the states. The FARS, in particular, being essentially a census, serves as a complete documentation of traffic fatalities in the nation.

Beginning in the late 1960's, the National Highway Traffic Safety Administration began to sponsor specialized field accident investigation teams under a multidisciplinary accident investigation (MDAI) program. While these teams initially reported detailed accident circumstances in several different forms, about 1969 the reporting became more consistent with adoption of the Collision Performance and Injury Report form. The MDAI data, as compiled for analysis, provided rich detail with regard to vehicle damage, injury to occupants, and causation, but--as discussed in the Statistical Inference report*--did not represent any defined population. Inferences drawn from the MDAI data have been necessarily guarded, and reference to other census-like data (e.g., police reports) has often been necessary to estimate occurrence frequencies in the accident population. The MDAI program fostered the development or improvement of many measurement methods which permit precise recording of information about accidents--notably the extensions of the Abbreviated Injury Scale, the Collision Deformation Scale, and methods for recording accident-causation information.

## 2.2 Capabilities of Present Systems

The questions given as examples above are indeed representative of the needs of the users. Most of the questions require two qualities in accident data--representativeness (so that inferences from the data can be drawn to a national or some identifiable large population), and detail (so that analyses of the data can

*Op.Cit.

produce accurate, reliable answers to particular questions). The need for completeness and precision in the data is implied.

The present data systems have been of limited usefulness to planners, evaluators, designers, and other decision-makers because (1) the police-reported data (or compilations of it, such as the National Accident Summary) do not provide the detail needed for solving particular problems; and (2) the MDAI-reported data do not well represent the national accident population. A further limitation of compilations of police-reported data is that reporting rules vary from state to state, and from one jurisdiction to another within a state. In injury reporting, for example, different meanings are assigned to the same scale--so that "A" injuries in one state are not strictly comparable in degree of severity with "A" injuries in another state. This subject is discussed in more detail in Volume II of a report by Scott and Carroll,* but it seems likely that precise and consistent reporting by all police agencies is not going to occur for many years. Blumenthal, et al., concluded that full national consistency of police accident reporting was not likely.**

MDAI data have the kind of detail necessary to determine meaningful injury severity distributions or injury mechanisms (to answer, for example, a question about the kinds and sources of injuries in side collisions). Indeed in their present format they have more detail than is necessary to answer the more common questions. These data have been used to provide insight into the need for standards, but most often they have been used as a basis for

*Acquisition of Information on Exposure and Non-Fatal Crashes, Volume II, Accident Data Inaccuracies, 1971, R. Scott, P. Carroll.

**A State Accident Investigation Program - Phase II, M. Blumenthal, H. Wuerdemann, July, 1969 (pg.36).

anecdotal examples rather than statistical interpretations. Arguments have been presented for such occupant-protection standards as limiting hood penetration of the windshield on the basis of a small number of observations of injuries caused by this mechanism, but the MDAI data cannot be used as a basis for a reliable estimate of the national incidence of these injuries. Imposition of such standards has been questioned on the basis of costs and benefits, and the data for computing such frequencies are not available.

The Fatal Accident Reporting System promises to be a most useful source of information. It has the advantage of being a census (i.e., a complete set of data) about fatal collisions. One disadvantage of a census is that one must accept whatever missing data comes about, and with limited funds it is difficult to reduce the missing data to a few percent. For the major factual data elements--driver age, sex, type of vehicle, time of day, etc.--the FARS provides a very good knowledge of fatal accidents. For less factual (or harder to get) elements, such as blood alcohol level, the time it took the ambulance to get to the scene, etc.--the percentage of missing data limits the usefulness of the data.

## 2.3 Summary

In summary, many decision-makers have questions which require precise and detailed information about the national accident population. And answers are needed for planning, countermeasure development, and designing. The data presently available have been used, but have been lacking in detail and representativeness: This report presents a new data acquisition and analysis system designed to overcome these shortcomings.

# 3. OVERVIEW OF THE NASS DESIGN

## 3.1 Objectives of a National Accident Sampling System

The NASS has three objectives:

(1) To create and maintain, by means of a sampling plan and diligent investigation, a set of accident data which truly represent the population of highway accidents in the United States, and which provide detailed information not regularly available in police reports. This aspect of the National Accident Sampling System will be referred to in this report as the Continuous Sampling Subsystem (CSS).

(2) To provide for acquisition, within the same population, of additional data about any chosen subset of accidents for finite time periods--in order to address special questions not answerable in the general data of the CSS. This aspect of the System will be referred to in this report as the Quick Reaction Subsystem (QRS).

(3) To provide a capability for in-depth investigation of selected accidents, so as to generate a set of very detailed reports pertinent to special problems. This aspect of the system is comparable to the existing in-depth investigation program of the NHTSA, and will be referred to in this report as the MDAI Subsystem.

## 3.2 Consideration of System Errors

In acquiring data to represent a population, errors or uncertainties arise from two broadly considered sources. Random error in an estimate of parameters measured within the data is a function of the sampling design and the number of cases investigated; and bias errors may arise because of inaccurate reporting, missing data (in which the desired parameters may have different values for the measured data), or inaccurate implementation of the sampling plan. These sources of error are pointed out to emphasize the necessity of considering both the mathematical system design and its operational implementation. Inaccurate reporting, inaccurate implementation of the sampling plan, and the failure to acquire specific data elements for a chosen report all contribute in essentially the same way to the bias term. They may all be thought of as data which should be there but are missing.

The random and bias types of errors, when estimating a proportion, may be represented as shown in Equation (1)

$$\underbrace{E_{RMS}}_{\substack{\text{the rms}\\ \text{error in}\\ \text{"p"}}} = \sqrt{\underbrace{\frac{pq}{N}\ (DEFF)}_{\substack{\text{Sampling}\\ \text{Variance}\\ \text{(random}\\ \text{error)}}} + \underbrace{\omega^2 \Delta p^2}_{(\text{bias})^2}} \tag{1}$$

where

$E_{RMS}$ = the range within which the true value will lie with a probability of about 2/3.

$p$ = the proportion being determined - e.g., the fraction of drivers wearing restraint systems.

$q$ = (1-p)

DEFF = the design effect (for a cluster design, generally a number greater than 1 which depends, among other things, on the areal variability of the quantity being measured).*

$\omega$ = the proportion of missing data.

*See Section 4 for a fuller definition of the design effect.

$\Delta p$ = the difference between "p" in the measured population and "p" in the missing data population.

The product pq is nearly invariant for values of p in the middle range. For example, for $p = 0.5$ and $q = 0.5$, $pq = 0.25$; and for $p = 0.3$ and $q = 0.7$, $pq = 0.21$. For very small values of p, pq is nearly equal to p.

In a sampling system with no missing data, the second term ($\omega$) would become zero, and the bounds on the error could be controlled simply by changing N and DEFF. But in any practical system there are always some missing data, and the second term must be considered. For example, if, in the acquisition of data, cases are chosen for inclusion in the sample on the basis of convenience (e.g., if a larger fraction of cases is obtained on weekdays than on weekends, because of the reduced work force on weekends), and if there is a difference in the proportion of people wearing restraints in the two periods, this will result in erroneous data. While values of DEFF can actually be determined only after the data are acquired, a sample problem with a typical and reasonable value will be presented here. If the design effect is 10, $p = 0.3$, and $N = 10,000$, the square root of the variance term is 0.014--thus the "one sigma" confidence interval on the estimate could be stated as 30% ± 1.4%. But with 25% missing data, and a true value of $p = .20$ in the missing data, the real error would be 2.8%. If the first term were reduced to zero, the error would still be 2.5%.

In many U-M ISR surveys, important differences have been found between persons who provide information on a particular variable and those who do not; it cannot be assumed that the information obtained from willing respondents is valid for persons who refuse to respond. In a survey conducted in driver licensing offices several years ago, persons who responded to a questionnaire had half as many convictions on their records as persons who did not.* In collecting accident data, for example, it may be easier to acquire information on serious accidents, and the missing data may contain a larger proportion of uninjured persons.

The percentage of missing data viewed as acceptable in a survey varies with the needs of the user. In a survey of drivers to determine the fraction of uninsured motorists in a state, the data to be used in connection with a no-fault constitutionality trial, Katz** went to great expense to have no missing data (detectives were used to track down "hard-to-find" individuals). But, in a more typical survey, such as the U.S. Bureau of Census truck transportation survey of 1972,*** there was approximately a 12% non-response. This was judged acceptable by the authors, although the instructions to users of the survey contain a warning that "our experience indicates that there are biases in the non-responding populations." In the present restraint system study being conducted by several accident investigation teams under MVMA and/or NHTSA sponsorship, efforts have been made to hold missing data down to 10%.

It is important in the design of a data collection system to consider both the sampling variance and the bias

*Exposure Study, Volume II, Carroll, op.cit.

**Katz, L. Presentation of a Confidence Interval Estimate as Evidence in a Legal Proceeding, July, 1974. Michigan State University, Department of Statistics and Probability.

***1972 Census of Transportation, U.S. Government Printing Office.

errors. The sampling error is mathematically tractable--i.e., it can be rather precisely defined as a function of the system design and the number of cases acquired. The bias error, on the other hand, might be characterized as insidious. For the most part, errors arising from the missing data are unknown, and critics of the output of the system are relatively free to suggest that the missing data values differ significantly from the known data values (always, of course, in a direction supportive of the critic's argument). While there are techniques for sampling of missing data cases to estimate bias, the preferred technique is to minimize the extent of the missing data in the operation of the system.

In the present report we devote considerable attention to both of these problems. The basic sample design starts with a list of U.S. counties and (for very small counties) county groups as a sampling frame, and defines a clustered controlled probability sample of these units based on both population and a travel-related parameter (gasoline sales). From this operation, several primary sampling units are defined in such a way as to represent all of the contiguous 48 states. Within primary sampling units, secondary and case sampling procedures are defined.

At the same time, appropriate attention has been given to the problems of achieving the necessary quality in the data to minimize the bias term. Experience with the restraint system program has been invaluable in this effort, and the system as presented here has many checks to ensure that data are properly reported. The self-checking capacity of the system is also described in detail.

Most of the error computations given in the remainder of the report do not include the bias term. The statistical design of the system is presented primarily in Sections 4 and 5, and the error statements there assume that the data are essentially unbiased. Sections 6 through 11 have to do mainly with system operation and implementation. The methods developed there are intended to minimize error due to missing or poorly reported data.

### 3.3 Major System Policy Choices

In defining a national system of accident data collection, it is necessary to place some bounds at the outset. While these may be changed later, the planning process cannot proceed very far without some specific guidelines. In this development program we have considered the problems of what types of accidents are to be represented, which might be intentionally excluded, what total population must be inferred to, and whether special emphasis should be placed on particular kinds of accidents. For the most part these become major policy choices which must be made by the agency responsible for the system, although there may be technical inputs which allow these questions to be resolved more easily. These design choices are discussed here.

Should the data resulting from this sampling system be capable of representing only the nation as a whole, or should it represent major subdivisions or even states? It has been assumed that the primary purpose of the National Accident Sampling System is to provide estimates for accident parameters for the nation. Using a cluster sampling technique, as discussed in Sections 4 and 5, it would be necessary to have a considerable number of primary sampling units in each sub-region to be able to obtain any meaningful measures about that area--certainly no less than eight. To represent the states individually, then, would require nearly 400 sampling units, and that is considered an unreasonable extension of a system aimed primarily at national statistics. In the design, however, it is possible to balance sampling units in as many as four regions of the United States, and thus to be able to draw some inferences about regional differences at that level. The system choice, then, has been to consider national representation as a primary factor, to attempt to achieve some capability to represent smaller regions, but to make no attempt to represent individual states. The purpose of NASS is to permit estimates of national statistics for national planning. It would take about as many primary sampling units to represent one state as it does to represent the U.S. But that is well beyond the projected scope and costs for NASS.

What accidents are to be represented? Traffic accidents range in severity from simple scratches of a fender to multiple fatalities. In attempting to estimate the national consequences of all traffic mishaps it would perhaps be useful to know about all of these. But a major fraction of the total dollar damage or total injury (weighted by severity) occurs in a more limited class of accidents. Candidates for such a limited class might be "all injury accidents," "all towaway accidents" (i.e., crashes from which at least one vehicle had to be towed from the scene), "all crashes in which at least one person was transported to a hospital," etc. The choice should satisfy two conditions: the resulting data should be useful in solving the problems the system is faced with, and the choice should be implementable in a consistent fashion across any of the jurisdictions which may be chosen for inclusion. This matter is discussed in some detail in Appendix A, but the choice for guiding further design has been to take at least fatal accidents plus all accidents from which at least one vehicle has been towed from the scene. At this time those sets seem to have the least variability from one region to another, and they contain nearly all of the serious injuries and most of the reported injuries.

Three other categories of accidents are of concern because they produce a substantial proportion of the nation's injuries. These are pedestrian, bicycle, and motorcycle accidents. The total set of accidents, then, to be used as candidates for investigation includes crashes in which a fatal injury occurred or from which at least one vehicle was towed from the scene, plus any other police-reported bicycle, motorcycle, or pedestrian accidents.

Should any regions of the U.S. be excluded from the sampling program? In the initial development of the sampling plan, consideration was given to including data only from Standard Metropolitan Statistical Areas (SMSA's)*--largely as a matter of convenience in collection of a set of data that would represent about 70% of the country. But it was ultimately concluded that representation of rural areas was important to develop national estimates of many parameters, so the present design takes into account all regions of the contiguous 48 states and the District of Columbia. Alaska and Hawaii have been excluded from the sampling frame at this time, although it would be possible to add them later to represent themselves.

Should there be an emphasis (within the chosen sampled accident population) on any particular kinds of accidents? Generally the answer to this question is yes. Experience in the restraint system study**(as well as other considerations) has led to the conclusion that there is more

*This was the basis for the preliminary plan discussed in Statistical Inference Report as well as in the contract work statement for this program.

**"A Sampling Program for Evaluation of the 1974 Restraint Systems", R. Scott and J. O'Day, SAE Paper No. 750188, February 24-28, 1975.

interest in injury-related questions, and that it will be more economical to emphasize the more severe accidents (by weighted sampling). Consequently there is provision in the design for weighting the selection of cases according to some prescribed rules.

What should the system operating costs be? Cost, of course, interacts with the other parameters of the system--accuracy, number of cases to be acquired, level of detail needed in the investigation, etc. In the same sense as in a set of algebraic equations, it should be possible to determine cost if all the other parameters are known, or to determine other parameters if cost is known. Based on discussions with the sponsor and a knowledge of operating costs in the present MDAI program, an annual operating cost of $6,000,000 to $12,000,000 has been assumed as a guideline for the system under discussion here. In Section 4, alternatives in design are discussed as a function of cost within this range.

Recommended choices then include aiming the system toward national representation; defining the population to be measured as fatal or towaway accidents plus pedestrian, bicycle, and motorcycle accidents; not including Alaska and Hawaii at the outset; planning for weighted case sampling to emphasize data of most interest; and designing toward an operating cost of $6 to $12 million per year.

### 3.4 Sampling Design

In developing a sample of accidents to represent the nation, sampling will be accomplished in as many as three stages. Primary sampling is the process of choosing sampling units from a frame of all counties (including groups of small counties) in the country. While the variance of the estimates made from the data will be sensitive to the number of primary sampling units, it will be a function also of the data itself. This interaction is discussed in Section 4 of this report. The design choice at this time is to take a sample of 35 primary sampling units--three densely populated areas (Los Angeles, Chicago, and New York) chosen with certainty, and the 32 others selected on a probabilistic basis but distributed equally in the four remaining equal population regions of the U.S.

Secondary sampling will be necessary in the primary sampling units that have a population greater than about 200,000 persons (i.e., about 3,000 accidents per year which would qualify as candidates for investigation by being fatals, "towaways," pedestrian, motorcycle, or bicycle. These 3,000 accidents can be considered to make up the local sampling "frame". It is important that this secondary sampling process represent the jurisdiction properly, and the method recommended is a two-step random sampling of sub-divisions and cases within the jurisdiction. This technique will approximate simple random sampling for the secondary process, and will minimize the variance as compared with other sampling methods. This is discussed in detail in Section 4.3.

Weighted sampling is introduced at case selection level. Given a sampling frame of, for example, 3,000 accidents within a Primary Sampling Unit, detailed investigations will be conducted of approximately one-sixth of these accidents. For estimating some simple parameters--e.g., the percentage of female drivers--variance would be minimized by drawing a simple random sample from the frame. For comparing injury-related factors (e.g., the number of drivers injured above and below AIS level "2" in small as opposed to large cars), the variance of the estimate can be improved by having more nearly equal populations in the compared injury groups. Such an improvement may be accomplished by taking a larger (but still random) fraction of cases in the more severe crashes, and a lower fraction in the less severe crashes--still reporting in detail on approximately 500 cases within a PSU.

For a particular problem, such as the one posed above regarding small and large cars and injury, it may be possible to compute an optimum weighting scheme based on historical data. The present system is designed to permit investigation of many different accident-related problems, and it is not reasonable to compute an optimum weighting scheme for all possible problems jointly. Anticipating that there will be greater interest in injury accidents involving relatively new-model vehicles, the tentative weighting assignment proposed here calls for the inclusion of all fatal accidents in the frame, along with a substantial proportion (say 50%) of the serious injury accidents involving new model vehicles, and lesser percentages of accidents involving older cars and light- or no-injury crashes. Reconstitution of the data by multiplying by the inverse of the sampling fractions is straightforward, although the variance computations are somewhat more complex (See Section 4.4). This weighted sampling procedure was used by most of the teams during the restraint system evaluation study, and the technique was quite useful there. This is discussed in more detail in Section 10, and an example of the weighted output is shown in Table 10.1, page 136.

## 3.5 Exposure Data and Other Considerations

When a change is recognized in the accident data, one is often faced with a problem of explaining why it occurred. If, for example, the average age of fatal accident drivers drops from 37 to 35* years over a period of several years, one could infer that (1) there are more young drivers now (or fewer old drivers) than there used to be, (2) young drivers drive more dangerously (or old drivers drive less dangerously) than they used to, or (3) young drivers are buying smaller cars which have a greater potential for injury to their occupants, etc. Exposure information is generally thought of as a second set of information bout the population at hand which will get the analyst one step closer to a solution.

In its simplest form, exposure information is already available, although not often with the precision necessary to a full solution. Male drivers have more accidents than females--about 70% to 30%. But it has been determined from several surveys that males do about 70% of the driving, and at least a first-order correction of the inference that males are poorer drivers can be made.

Occasional data are available which show, for example, truck and car mileage by time of day, and national and state estimates of total mileage are made from analyses of gasoline sales. But there is presently not a national set of exposure information which could adequately serve to permit adjustment of the accident data set defined in this study. By definition the present project was not expected to solve the "exposure" problem, but a few comments on it are in order.

In early discussions of a national accident data collection system one of the techniques considered was a general (e.g., a household) survey. Although such surveys have been made for the purpose of counting

*See Figure 10.1, Section 10, page 137.

numbers of accidents, the level of detail of crash and injury information indicated by the questions at the beginning of this section could not be obtained in that way. Further, it has been demonstrated that people tend to forget even severe accidents rather quickly, and that a maximum of 90 days is the most cost-effective recall period for such a survey.* A general population survey would be a very useful supplement to the present sampling plan as a means of obtaining counts of types of accidents not completely reported in police accident files.

Exposure information also might be obtained by such a direct survey of the general population. While only a small percentage of the populace has accidents, nearly everyone has some exposure to the highways. It is suggested that some sort of general population survey--by household interview, random digit dialing by telephone, interviews of drivers at license renewal time, etc.--could be accomplished within the same primary sampling units defined for the accident investigation program.

The system is intended to represent a certain class of collisions as defined above--namely, those accidents from which at least one vehicle was towed and all police-reported fatal and pedestrian-bicycle-motorcycle accidents. Towaway is expected to be an implementable criterion in most jurisdictions because (1) most police reports indicate disposition of the vehicles in sufficient detail to identify those which have been towed from the scene, and (2) Highway Safety Standard #18 sets a goal for police agencies to make written reports on all towaway accidents. In a formal check of this

*National Center for Health Statistics, HEW, Optimum Recall Period for Reporting Persons Injured in Motor Vehicle Accidents, Washington, D.C. US GPO, April, 1972 (Vital and Health Statistics - Series 2 - No.50).

expectation a survey of towing agencies and insurance companies was compared with police-reported accidents in one county in Michigan, and it was concluded that fewer than 5% of the reports in the first two were not available in the police record. It is likely that this will not be universally the case--at least at the beginning of system operation. This sort of survey should be conducted in each chosen primary sampling unit to determine whether there will be many missing cases. Alternatives, including substitution of primary sampling units, or construction of a multiple frame (i.e., where the towaway population is defined by looking at both the police reports and some other set of reports such as those of towing agencies), are possible. This is discussed in Section 4.2.

## 3.6 Data Handling

Although the proposed data-handling method results in the relatively rapid construction of a representative set of data, speed was not the original purpose of the design. Experience with past programs, however, has indicated that sequences of paper-handling result in relatively long delays, and that when there is need for a follow-up investigation to obtain some missing data, the time has often passed when that can be accomplished. Present experience with MDAI reports shows an average of 10% of the data elements are incompletely reported for each case, and most cases edited in the MDAI process exhibit some inconsistencies in reporting. These vary from completely missing reports on occupants (e.g., three occupants are reported in the summary of the case, but only two occupant reports are completed) to a fender reported damaged in one place in the report and undamaged in the other. Similar though perhaps less sophisticated inconsistencies occur regularly in police reports--e.g., with a VIN number indicating a truck and a report stating that the vehicle was a passenger car. In the MDAI cases every effort is made to resolve inconsistencies before placing data into a digital file, but it is often not possible to go back to the original source of the data because of the passage of time. The rapid handling described here, then, is mainly to ensure completeness and accuracy, and to produce evidence of discrepancies soon enough so that they can be resolved.

## 3.7 Physical and Operational Description of the System

The suggested framework for the National Accident Sampling System is a sample of 35 Primary Sampling Units distributed to represent four major regions of the United States. Within each PSU a local sampling frame of at least 400 fatal-towaway-pedestrian-bicycle-motorcycle crashes is defined, and a more detailed investigation and reporting is performed on a subset of these.

The 35 Primary Sampling Units will be divided into five or six groups, each managed by a zone* control unit. Each of the zone control units will serve several functions--assisting in technical problems at any of its primary sampling units, aiding in initial setup of relations with local police, hospitals, etc., and providing the full in-depth (MDAI) capability for that region of the country. It is possible for the zone control unit to be geographically located at the same place as a primary sampling unit, but it is not necessary.

*Zones are adminstrative units as opposed to regions discussed earlier as sampling areas. See Section 11.

Data for the local frame are derived generally from police reports, and are reported frequently (probably daily) by a digital communications system to the system data center at the national level. Selection of cases from the frame for more detailed investigation (i.e., for the CSS) may be made by a weighted sampling technique monitored at the national data center, and the detailed investigation reports are subsequently reported to the center in the same manner.

Feedback from the data center to the sampling unit is accomplished largely through the same communications network, but, for purposes of management, several interimediate centers are placed in the loop. The lines of communication and control are discussed in Section 11. Figure 11-4 shows the basic arrangement of a system with many primary sampling units and a few zone control centers.

### 3.8 Functions

The system is designed to meet the requirements of a continuous sampling system (CSS). For this aspect of the system, each primary sampling unit is staffed with a team of four to six persons who are responsible for continuous collection of the frame data, inputting of that information to the data center, investigating in detail the assigned accidents, and inputting the resulting information to the national sample file.

At the national level the "frame" file becomes a set of at least 100,000 accidents per year with a limited amount of data--date and time of the accident, severity (using a police injury scale), makes and models of vehicles, and certain other data required for the weighted random selection rules to be applied. It is expected that the frame file would be substantially complete within a week or less of the occurrence of the crash, and entries would be 90% complete within 48 hours.

The detailed investigation file at the national level contains much more information about the crash and each of the vehicles and persons involved. It is constructed automatically from the inputted reports, and is periodically divided into three working analysis files centered on the accident, the vehicle, and the people. It is expected that entries into this file would normally be made within a few days, but that some cases might require more time than this. It should be essentially complete within a month of the accident occurrence.

While the basic design is centered on the CSS portion of the system, the QRS portion follows the same routines. The computer programs generated for creating the interactive input routines are general in nature--i.e., it is possible to change the requested data elements in a relatively simple manner. Thus, a new query system can be implemented within hours after the questions are defined, and the responses to that query will be automatically built into a working file for analysis of a particular problem. This is discussed in more detail in Section 12.

The full MDAI in-depth cases are not at this time planned to be part of a statistically defined sample of accidents. In-depth investigations have been proven to be useful in gaining insight into particular subjects--e.g., school bus accidents, air bag cars, motor homes, etc. They tend to provide a level of detail not available in the CSS or QRS parts of the system, and would be conducted in the present design only by teams operating out of the zone control centers. Such teams would have in them or available to them a level of expertise in vehicle factors, human factors, and environmental factors appropriate to the in-depth investigations, and their resultant data would be handled in the same way that past MDAI cases have been.* In addition, these teams would be available for special assignment to accidents of interest to the Department of Transportation as necessary.

## 3.9 Evolution of the System

The National Accident Sampling System is intended to be a continuous activity, supplying information to the federal government and to the nation about highway traffic accidents. It has been argued here that 35 primary sampling units distributed throughout the country will, given collection of the number of cases prescribed, provide an adequate representation of the country. There are, however, several system choices which may be modified before such a system is fully operational.

The present restraint system study** has served as a sort of pre-pilot operation of a sampling system, and its execution in several areas of the country has brought to attention regional differences in the form of police data and in methods of interacting with local officials. Data handling in the restraint system study has many parallels with the NASS design, although the regions were not selected on a probability sampling basis and strong inferences to the national population are not possible.

*Presently MDAI cases are coded into digital form and stored at The University of Michigan, where they can be addressed from remote terminals by NHTSA and other agencies.

**O'Day, Scott, Op.Cit.

This report comes at the end of the NASS design phase. The most appropriate next step would be a pilot operation intended to further define such choices as (1) digital-interactive vs. paper reporting by teams, (2) the appropriate number of PSUs to be managed by each zone control center, (3) final definition of team staffing requirements, (4) operation of a Quick Reaction System problem with more than one PSU, etc. These are discussed in Section 11 and 12 in more detail.

While it would be possible to proceed to 35 primary sampling units over a short period of time, that course is not recommended. Many system details will be better defined with experience, and a progression through several pilot phases is recommended. NHTSA should implement a pilot operation of the system as quickly as possible at one location, and over a period of one year implement four additional PSUs, each under the control of a different zone control center. While one year of experience with four or five PSUs is not likely to provide defensible national statistics, the next recommended expansion to about 16 in a second year would. An expansion to 35 PSUs could be accomplished during a third year. The procedure for drawing the sample of PSUs discussed in Section 5 permits a phased expansion of operational teams in a nationally representative manner up to 67 sites.

The choice of financial structure for operating the NASS is really beyond the concerns of this design project. The final system should be viewed as a continuing federal effort, and it may be that it could be managed most conveniently within the civil service system. Zone control centers could be located at or in conjunction with DOT regional offices; the computer and data center could reside in the NHTSA offices in Washington; and the local PSU personnel could be DOT employees residing in their assigned areas. Alternatively, the bulk of the system could be contracted by NHTSA in parts, with appropriate bidding procedures and choice of contractors based on expected performance and cost. In this latter case there would be a strong argument for relatively long-term contracts--perhaps three to five years--and with various contractors phased to renew at different times so as to minimize perturbations to the system.

For the development and pilot testing programs it is deemed appropriate to make use of existing capabilities in the MDAI field activities throughout the country. Note that a strict probability sample is not likely to select a geographic location for a primary sampling unit which is coincident with a present MDAI team operation, with the exception of teams located in the very largest counties which will be included in the sample with probability equal to one. But present MDAI teams have established remote investigation units at some distance from their home base, and could be expected to manage subsidiary Primary Sampling Units within a region of the country. We recommend that, for the pilot phase of this program, advantage be taken of the experience of several local teams by using them in this way.

# 4. THE SAMPLE DESIGN

The goal of the sample design is to obtain a representative national sample of fatal, towaway, pedestrian, bicycle, and motorcycle accidents (as discussed in Section 3). This sample inevitably must be a statistical compromise between maximizing the statistical reliability of the data analysis and minimizing the costs of the data collection. While a national simple random sample of all the accident types listed above would probably provide much higher precision of analytical findings, this approach was rejected for several practical reasons. In the first place, there is no single national list of such accidents from which to draw a sample. Even if the various centralized state accident files were utilized for this purpose, there would be the problem of incompleteness in many states and the problem of long lapses between the time of the selected accident and its investigation which would result in too much incomplete and unreliable data. Secondly, there is the logistical problem that sending accident investigators to all parts of the United States would result in very high costs per accident investigated. It is obvious that an accident investigation team would be able to investigate a great many more accidents for the same cost if it were located in a limited geographic area than if the team members had to travel to all parts of a state or larger region to investigate a simple random sample of accidents.

Therefore, this report proposes a clustered controlled probability sample design in which the selected accidents are clustered by specific geographic areas but in which stratification is used in the selection of these areas in order to ensure diversity in the types of accidents investigated.

When a sample survey is first designed in a new area such as accident data collection, one typically has far from perfect knowledge about parameters of cost, of variance, and of the most important statistics the survey must produce. Hence the statistical compromise cannot be optimal in any time sense. However, from the survey experience itself one can obtain the data needed for later improvments through evolutionary development based on methodological research.

## 4.1 The First Stage: Choosing the Primary Sampling Units

Several interactive issues had to be considered in developing a method for choosing a sample of primary sampling units (PSUs). These include what type or types of PSUs to utilize, what measure of size to use in the selection process, what stratification variables to include, and how many PSUs to select with how many expected accidents to investigate per PSU.

The last issue involves the questions of average costs per investigated accident and the relative levels of precision in statistical estimates to be expected from clustering different numbers of investigated accidents in different numbers of primary sampling units. If one could assume that all potential PSUs were completely uniform in the types of accidents which take place within their borders, then obviously the most cost-effective approach would be to select one large PSU with one large accident investigation team, and the sampling error would be a straightforward calculation based on the square root of the total sample size in that PSU (SampEr = $2\sqrt{pq/N}$).* However, in the real world there are substantial differences in the characteristics of accidents among

*In this section of the report the sampling error is defined as 2-sigma, that is the range to give a 95% confidence interval.

different geographic areas--for example, rear-end collisions occur less frequently in rural areas than in urban areas.

To explore how much this variability might affect the precision of sample estimates, several variables of the type commonly available in police reports were analyzed in two HSRI accident files. These files contain five percent samples of all 1973 police-reported accidents in Texas and Michigan. In Michigan the tabulations were based on dividing the state into 34 potential PSUs which were either whole counties or groups of counties with a minimum 1970 population of about 85,000. In Texas the tabulations were based on dividing the state into 56 potential PSUs which were either whole counties or groups of counties with a minimum 1970 population of about 98,000. In both states the PSUs were divided into three strata: one composed primarily of central counties of Standard Metropolitan Statistical Areas (SMSAs), one composed primarily of suburban SMSA counties, and one composed primarily of non-SMSA (i.e., rural) counties. The variance terms were calculated separately for each of the three strata, and then the overall variance was calculated by the formula:

$$\mathrm{Var}(\frac{y}{x}) = \frac{1}{(x)^2}[\sum^{3} \mathrm{Var}(y_n) + r^2 \sum^{3} \mathrm{Var}(x_n)$$

$$- 2r \sum^{3} \mathrm{cov}(y_n x_n)]$$

where:

r = the overall proportion on some characteristic.

y = the number of cases with this characteristic.

x = the number of non-missing-data cases.

n = the number of PSUs in a stratum.

The Michigan results are based on 7,490 vehicles (involved in at least 4,721 accidents) which were towed from the scene (about 25% of the total traffic unit sample). The Texas results are based on 9,498 vehicles (involving at least 6,000 accidents) which were assigned a score of 3-7 on the vehicle damage scale (about 30% of the non-missing-data sample; but 19% of the Texas traffic units were not scored on this variable).*

The results of these calculations are presented in Tables 4.1 and 4.2. Shown for each variable are its proportion of the total N, its sampling error (two times the standard error, thus using a 95% level of confidence), its cluster design effect (DEFF is the ratio between the actual variance and the simple random sample variance for an N of the same size), the square root of DEFF (which is the ratio of the actual sampling error to the simple random sample sampling error for an N of equal size), its rate of homogeneity** (roh = $\frac{\text{DEFF-1}}{\text{b-1}}$ where b is the average number of cases per PSU), and the percentage error of the obtained proportion (the sampling error divided by the proportion).

It can be quickly seen that there is substantial heterogeneity among the PSUs in both states on some of the variables of interest (e.g., the proportion of single vehicle crashes), while some other variables (e.g., Sunday crashes) tend to be quite homogenously distributed throughout each state. These data from all counties in two states cannot provide a perfect representation of what would be expected in a national sample of PSUs, but at

*"Towaway" is not a coded variable in the Texas data, and scores of 3-7 on the vehicle damage scale (also called TAD) were used as a surrogate for towaway.

**This is sometimes called the intra-cluster correlation coefficient. A roh of less than .005 indicates very little variability among clusters, while a roh approaching .1 indicates great variability among clusters.

TABLE 4.1. SAMPLE STATISTICS FOR SELECTED VARIABLES IN THE 1973 TEXAS ACCIDENT FILE (A FIVE PERCENT SAMPLE), USING VEHICLES SCORED IN THE 3-7 RANGE ON THE DAMAGE SCALE.

N=9,498

| Variable | Percent | Sampling Error | Design Effect | $\sqrt{\text{DEFF}}$ | roh | Percentage Error |
|---|---|---|---|---|---|---|
| Interstate Highway Accident | 11.3 | 2.2 | 7.4 | 2.72 | .0596 | 19.5 |
| Single Vehicle Accident | 38.0 | 3.2 | 6.4 | 2.53 | .0503 | 8.4 |
| Accident Involving a Ped/Bic | .12 | .08 | .9 | .949 | -.0008 | 66.7 |
| Fatal Accident | 1.96 | .49 | 2.0 | 1.41 | .0089 | 25.0 |
| Nighttime Accident | 36.9 | 1.4 | 1.2 | 1.10 | .0021 | 3.8 |
| Sunday Accident | 13.5 | 1.2 | 1.9 | 1.38 | .0085 | 8.9 |
| Accident With an "A" Injury | 10.0 | 1.2 | 2.3 | 1.52 | .0124 | 12.0 |
| Accident with No Injuries | 58.1 | 1.8 | 2.0 | 1.41 | .0095 | 3.1 |
| Pre-1972 Model Vehicle | 76.3 | 1.1 | 1.6 | 1.26 | .0033 | 1.4 |
| Driver Under Age 65 | 94.0 | .68 | 2.0 | 1.41 | .0057 | 0.7 |
| Truck Vehicle Type | 14.4 | 1.9 | 6.9 | 2.63 | .0349 | 13.2 |
| Female Driver | 28.7 | 1.6 | 2.9 | 1.70 | .0111 | 5.6 |
| "C" Injury to a Vehicle Occupant | 9.5 | .85 | 2.0 | 1.41 | .0059 | 8.9 |

TABLE 4.2. SAMPLE STATISTICS FOR SELECTED VARIABLES IN THE 1973 MICHIGAN ACCIDENT FILE (A FIVE PERCENT SAMPLE), USING VEHICLES TOWED FROM THE ACCIDENT SCENE

N=7,490

| Variable | Percent | Sampling Error | Design Effect | $\sqrt{DEFF}$ | roh | Percentage Error |
|---|---|---|---|---|---|---|
| Interstate Highway Accident | 6.4 | 2.1 | 9.1 | 3.02 | .059 | 32.8 |
| Single Vehicle Accident | 47.3 | 5.7 | 15.3 | 3.91 | .104 | 12.1 |
| Accident Involving a Ped/Bic | .49 | .21 | 1.1 | 1.05 | .001 | 42.9 |
| Fatal Accident | 1.7 | .64 | 3.0 | 1.73 | .114 | 37.6 |
| Nighttime Accident | 43.0 | 1.2 | 0.7 | 0.84 | -.002 | 2.8 |
| Sunday Accident | 14.0 | 1.0 | 1.0 | 1.00 | .001 | 7.1 |
| Accident with an "A" Injury | 12.6 | 1.3 | 1.7 | 1.30 | .005 | 10.3 |
| Accident with No Injuries | 45.0 | 3.2 | 4.7 | 2.17 | .027 | 7.1 |
| Pre-1972 Model Vehicle | 74.0 | 1.5 | 2.1 | 1.45 | .005 | 2.0 |
| Driver Under Age 65 | 93.9 | 1.4 | 6.0 | 2.45 | .023 | 1.5 |
| Truck Vehicle Type | 8.8 | 2.3 | 12.6 | 3.55 | .053 | 26.1 |
| Female Driver | 27.1 | 2.1 | 4.1 | 2.02 | .014 | 7.7 |
| "C" Injury to a Vehicle Occupant | 19.0 | 3.0 | 11.2 | 3.35 | .047 | 15.8 |
| Motorcycle Vehicle Type | 2.3 | 0.3 | 0.7 | 0.84 | -.001 | 13.0 |

least they are sufficient to demonstrate some relative expected magnitudes of clustering effect on the sample estimates for a selection of variables, and they do indicate that for many variables the variance in a national clustered sample is expected to be substantially greater than it would be in a national simple random sample. Nevertheless there seems to be no practical alternative to the clustered design approach with a resident accident investigation team in each cluster which is able to collect complete detailed information on the selected accidents in a cost-efficient and timely manner. Of course the larger numbers of the NASS sample would provide lower sampling errors than those shown here, but the design effects shown here may be fairly typical of what can be expected in the NASS sample.

In order to estimate an optimum number of sample areas in which to establish investigation teams, the Texas and Michigan data on rates of homogeneity were used to estimate the sampling errors of several variables with different numbers of PSUs and average Ns per cluster. First, however, in order to set some reasonable limits on this process, estimates were developed of the relative costs of operating various numbers of teams with varying numbers of accidents to be investigated per team. Two sets of error curves were then calculated with the accident data from each state, one utilizing an estimated overall operational cost of $6,000,000 and the other using an estimated overall cost of $12,000,000. Obviously, other sets of curves for different total costs could be calculated, but these costs bound the range suggested by the sponsor for planning purposes, and a perusal of these error curves will provide some idea of the relative precision of sample estimates to be expected for various types of variables at these two magnitudes of effort.

The error curves are presented in Figures 4.1 to 4.4. As would be expected, the optimum combination of number of PSUs and average PSU "take" varies considerably among the different variables. Since there is no one variable which can be considered controlling in the choice of the optimum combination, some judgment must still be exercised in recommending the "best" sample sizes at the two different cost levels. At the $6,000,000 cost figure a sample of 32-36 PSUs averaging about 350-505 investigated accidents per PSU would seem optimum in terms of relative levels of precision of sample estimates and in terms of providing enough work to keep full-time investigating teams appropriately occupied. About twice this number of PSUs, 64-72, with similar average PSU "takes" would seem optimum at the $12,000,000 level of effort.

Thus the total sample size would be 12,000-16,160 accidents involving about 20,000-26,000 traffic units at the $6,000,000 level and 25,920-35,200 accidents involving about 41,000-56,000 traffic units at the $12,000,000 level.

Having determined that a reasonable level of effort for each accident investigation team would be 350-550 accidents per year, the next issue to be settled concerns establishing a minimum PSU size to generate enough accidents which meet the selection criteria. Again the Texas and Michigan 5% sample files for 1973 were used to see how the various types of accidents of interest were distributed geographically among the 56 and 34 regions in the two states.

In Texas the smallest number of 1973 seriously damaged vehicles (3-7 on the damage scale) was found in a sparsely populated western area containing 16 counties and

99,000 people. This number was 41, or a projection of 820 when multiplied by 20 to account for the 5% sample used. The second lowest number, 44, was found in the two counties around Abilene with a population of 116,000. These two geographic areas were not the lowest in reported pedestrian-bicyclist-motorcyclist accidents, but only four areas had fewer such accidents than the three included in the 5% sample from the first area mentioned above. Thus, assuming that 20% of the vehicles not classified on vehicle damage were also really 3-7 on the damage scale but that each 3-7 damaged vehicle represents only .8 accidents (because some of the time two such vehicles are in the same accident), one obtains in Texas a minimum projection of 748 accidents of the types to be investigated in a potential PSU of a minimum 100,000 population.

In Michigan the lowest number of towaway vehicles in the 5% sample was found in a rural area in the western Upper Peninsula, containing six counties and 91,000 people. This number was 39, which projects to 780 when multiplied by 20. The second lowest number, 49, was found in an adjacent area of three Upper Peninsula counties containing 85,000 people. Pedestrian-bicyclist-motocyclist accidents are somewhat more frequent in the Michigan file (5.6%) than in the Texas file (3.9%), and even the smallest Michigan regions have at least five such accidents in the file. It seems reasonable to project in Michigan a minimum of 724 relevant accidents in a potential PSU of a minimum 85,000 population, taking into account some duplication of towaway vehicles in the same accidents.

It should be emphasized that these low numbers of police-reported accidents were not typical in either Michigan or Texas. Most of the small PSUs had considerably more accidents in the files. The number of accidents involving vehicles scored 3-7 on the damage scale per 100,000 population in the 5% Texas file was over 80, which projects to 1,600 per 100,000 people; and the number of towaway accidents per 100,000 population in the Michigan 5% sample was slightly greater.

On the basis of these Texas and Michigan distributions it was decided that an accident investigation team should serve an area with a resident population of at least 50,000 people in order to be fairly certain of having a reasonable number of accidents to investigate (i.e., enough to keep a team occupied). Even with the reduction in the incidence of traffic accidents since 1973, almost any geographic area of this minimum size could be expected to provide at least 350 accidents per year of the types to be investigated. Most such minimum-sized areas would offer more accidents than one team could handle, thus permitting a further selection process to better represent accident types of greatest interest.

In regard to type of geographic area to use as a basis for the primary sampling, it is clear that most states would be too large for one team to cover efficiently, while the vast majority of the over 12,000 police agencies in the United State would be too small to serve a complete accident investigation team. A geographic unit which aggregates a number of police agencies is needed, and for this purpose the county seems to be an ideal unit. Logistically most counties are small enough in area to provide a reasonably bounded working area for a resident accident investigation team, and yet counties also tend to be large enough to provide considerable heterogeneity in accident types, a significant advantage in terms of minimizing sampling errors.

About 80% of the U.S. population live in counties of at least 50,000 population. Unfortunately, however, nearly 80% of the 3,108 counties, parishes, and independent cities in the contiguous United States are less than 50,000 in population, ranging down to the 95 persons estimated for Loving County, Texas, in 1973. Therefore it is necessary to group these smaller counties into multi-county potential PSUs with a minimum population of 50,000. In some sparsely settled areas this will result in some very extensive potential PSUs, and perhaps an accident investigation team would have to set up dispersed substations in such an area. However, only about one-fifth of the chosen PSUs are likely to be multi-county units, and most of these will contain no more than three to four counties.

A remaining issue is what type of measure of size to use in selecting the PSUs by the controlled probability-proportionate to-size procedures. Ideally, since the purpose of the sample is to draw a national sample of certain types of accidents, the most appropriate basis of the sample selection would be the number of accidents of these types taking place in each potential PSU. However, while most states publish some kinds of accident data by county, there are serious problems in the comparability of these data from state to state and even from county to county within the same state. An inquiry into published data from 14 states found 1971 per capita accident rates varying from .021 in New York to .050 in Colorado (not reported at all in California), while injury accident rates varied from .005 in South Carolina to .012 in New York. Thus the same state, New York, was lowest on one accident measure and highest on another. There undoubtedly are real differences in accident'rates among the various states, but it seems unlikely that these differences are accurately reflected in published state and county non-fatal accident statistics.

Other types of size measures which were considered because they seemed to have a relationship to accident rates were population, number of vehicle registrations, number of licensed drivers, miles of primary roads, gasoline service station retail sales, and eating and drinking place retail sales. However, only the two retail sales variables showed correlations with Michigan county accident totals as high as the accident-population correlations. In Michigan counties, for example, these three variables correlated with both total accidents and injury accidents at figures above .99.

Therefore it was decided to use county population as the basic measure of size for selecting the PSUs. However, it was recognized that there are many rural areas which have disproportionately high accident rates due to heavy visitor traffic, and it was decided that gasoline sales data should be used as a stratification variable to ensure that counties and county groups with diverse levels of gasoline sales are adequately represented in the national sample.

It was also decided to use two geographic variables as stratification factors in the selection of the primary sampling units. While it is doubtful that there are great differences in accident characteristics among the major regions of the United States, it seems important from at least a public relations point of view to ensure that there is adequate representation in the national sample from each major section of the country (Northeast, Midwest, South, and West). So these regions serve as the first level of stratification.

The third stratification variable is the rural-urban factor. There are significant differences in accident characteristics between urban and rural areas, so it is essential that the sample design provide for adequate representation of these differences. Accordingly, the potential PSUs are classified into three basic rural-urban types: central counties of Standard Metropolitan Statistical Areas (SMSAs), suburban counties of SMSAs, and non-SMSA counties. In 1973 about 73% of the United States population was residing in Census-classified SMSAs, and almost 55% were in the central counties of these SMSAs.

It is not known how much these three stratification variables will contribute to variance reduction for statistics based on the sample data. When the variance was calculated for selected Texas and Michigan accident variables both with and without an urbanization stratification factor the results were rather mixed. Some variables showed a higher variance with rural-urban stratification, while some other variables showed a lower variance with rural-urban stratification.

In summary, then, the sample design for the selection of the primary sampling areas involves the choice by a controlled probability-proportionate-to-population technique of 32-72 counties and county-groups each with a minimum population of 50,000 people. These selections will be made from within strata established by the intersection of three stratification variables: region of the United States, degree of urbanization, and per capita gasoline sales.

Within each selected PSU a resident accident investigation team will be established, and each team is expected to investigate between 350 and 550 accidents per year. Thus, depending on the total number of PSUs, a national sample of 12,000 to 35,000 accidents (involving 20,000-56,000 traffic units) would be produced. The actual number of PSUs to be selected depends primarily on the level of financial support which the government plans to give this enterprise. The suggested range represents an estimated overall cost varying from $6,000,000 to $12,000,000. The larger the number of PSUs selected, the greater would be the national sample size and the higher would be the precision in the statistical estimates derived from the sample.

## 4.2 Choosing the Sample of Accidents

It is expected that the accident records of all police agencies operating within the geographic boundaries of a selected PSU will serve as the basic source of the accidents to be investigated. Fortunately, local police agency jurisdictions are almost always circumscribed by county boundaries except for the few towns and cities which are located in two counties. However, because this is not true for state police operations, in all selected PSUs in which state police or highway patrol personnel sometimes make accident reports it will be necessary to use the accident files collected at the state stations in or near the PSU county or counties in order to obtain information on accidents in that PSU which were reported only to state authorities. While it sometimes happens that police personnel from two different agencies will provide assistance at one accident, duplicate reports of the same accident are not likely to be made very often. However, the possibility of such duplication will be checked in the accident investigation team's accident screening process.

In all but the very smallest PSUs the number of police agency records of fatal, towaway, pedestrian, bicyclist, or motorcyclist accidents is expected to be greater than the 350-550 accidents which each team will investigate each year. Therefore, a controlled probability procedure must be developed for choosing which accidents to investigate. This involves a two-step process. The first step will be to go through the accident files listing basic information about all accidents which meet the designated selection criteria (fatal, towaway, etc.). This list will then serve as the PSU accident frame from which particular accidents will be selected for inclusion in the accident investigation sample. The basic information to be listed for each accident from the police records would include such items as the type of accident, the type of roadway, injury severity, time and date, vehicle types and model years, driver characteristics, the accident record number, and perhaps vehicle license numbers and/or VIN numbers.

One of the first tasks in each PSU will be to look through whatever accident summaries are already available from the various police agencies in order to estimate the total numbers of the designated types of accidents which usually take place within the PSU. It may be that insufficient information of this sort will be already available, and it may be necessary to actually go through the previous year's accident records in order to obtain sufficient estimates of these numbers. Once an estimate of the total number of accidents of the designated types is obtained, it is a simple matter to calculate the overall PSU sampling rate by dividing this number into the number of accidents to be investigated in that PSU. For example, if the estimate of the number of eligible accidents per year in a PSU was 2,500, and the team was to investigate 500 of them, the overall PSU sampling rate would be .20.

The simplest approach to choosing the PSU accident sample from the PSU accident list (frame) would be to follow a random procedure. For example, with a 20% sample rate two one-digit numbers (e.g., 4 and 0) might be chosen from a table of random numbers, and then all accidents whose accident record numbers ended in these two digits might be selected for investigation. This approach would also have the statistical advantage of providing for the lowest possible variance in the PSU sample of investigated accidents.

The disadvantage of this simple random sample approach to selecting PSU accidents for investigation is that a great deal of the accident investigation team's time would be spent on types of accidents which are not of great interest (e.g., involving older vehicles in which no one was injured), and only small numbers of accidents would be investigated of the types which would be of greatest interest to the NHTSA standards-setting personnel (e.g., fatal accidents involving late-model cars). Therefore, it is recommended that differential sampling fractions be used in each PSU for different types of accidents, with over-sampling of accidents involving greater severity and more recent model cars and undersampling of accidents involving less severity and older cars. While this approach would lead to somewhat increased sampling errors in analyses using the entire accident sample, it would greatly reduce the sampling errors in analyses of those subgroups of greates interest to accident researchers.

The particular sampling ratios to be used with the different types of accidents in a PSU depend both on further discussions with NHTSA personnel concerning the accident subclasses of greatest interest <u>and</u> on the types of accident classification information available in the PSU police records. For example, in the Oakland County and Washtenaw County restraint system studies, HSRI intentionally oversampled accidents involving a hospitalized participant. But if information about whether or not a victim was taken to a hospital is lacking in many PSU police accident records, then that variable would not be an appropriate one for categorizing accident subclasses with different sampling rates in that PSU. Fortunately, it is not essential that the same subclasses be used among all PSUs, since the weighting factors to take into account the differential sampling rates will be determined at the PSU level and will not affect the summarizing of all PSU data in the national sample.

An example of a possible system of differential sampling rates to be used in a hypothetical PSU with 2,500 annual accidents in its frame is given below:

| Accident Class | Frame N | Sample Rate | Sample N |
|---|---|---|---|
| Fatal | 40 | 100% | 40 |
| Late Model, Hospitalized | 170 | 50% | 85 |
| Late Model, Not Hospitalized | 370 | 20% | 74 |
| Early Model, Hospitalized | 520 | 20% | 104 |
| Early Model, Not Hospitalized | 1,100 | 10% | 110 |
| Pedestrian and Bicyclist | 150 | 30% | 45 |
| Motorcyclist | 150 | 30% | 45 |
| TOTAL | 2,500 | 20.1% | 503 |

Thus the procedure to be followed would involve team members (1) going through the 6,000 or so police accident records to pick out those accidents which meet the basic selection criteria; (2) listing basic information from these 2,500 eligible accidents; (3) classifying these eligible accidents into their appropriate subclasses; and (4) selecting certain accidents for investigation from each subclass according to a prescribed random method in relation to the PSU sampling rate for that subclass. This method would involve making use of particular predetermined digits of some number (or part of a number) on the accident record form which could be expected to vary from form to form in a random fashion--perhaps the police-assigned accident identification number or the license or VIN number of the first vehicle. It is important that this selection procedure be prescribed clearly in order to permit easy monitoring of the selection procedure and to prevent any staff member biases from affecting the selection process. The team member would be able to select the particular accidents for investigation at the time he is listing the accident frame, and he would thus be able to record all needed driver identification information, etc., for the selected accidents at the same time.

It is planned that the basic information collected on each accident in the frame will be entered into a computer file, and thus there will be a computerized method for checking that proper selection procedures were followed in regard to each sample accident, and non-sampled accidents which should have been selected will also be discovered. It would also be desirable to periodically check the accident records from the PSU in the state centralized accident files (if available) in order to ascertain if there are eligible accidents in the PSU which are not being entered in the PSU accident frame, either due to team member carelessness or due to deficiencies in the process by which police agencies make their records available to the team.

An alternative procedure to having the teams select accidents from the frame for detailed investigation, is to have the computer make the selection at the time the frame is entered. This is a highly desirable feature of the system, described in more detail in Section 7. The program to accomplish the selection, called RULE (see Appendix B), operates in its present version as a simple random selection based on a predetermined set of selection probabilities. In the future, it is contemplated that the selection would be made adaptive to account for variations in the accident production at specific locations.

There are two other important methods which should be used in checking the completeness of the PSU accident frame. The first involves conducting a general population sample survey in each PSU (probably by telephone, using random digit dialing techniques) to find out about residents' recent accident experience and whether any fatal, towaway, or pedestrian-bicyclist-motorcyclist accidents in which they or their family members were involved were not reported to the police. If an accident had been reported to the police in the PSU, enough information would be obtained from the respondent to enable checking whether or not that accident was included in the PSU accident frame. Information about eligible accidents which had not been reported to the police would permit developing estimates of the completeness of the different classes of accidents in the accident frame, and if sizable deficiencies in the frame were discovered these estimates could even be used for weighting the sampled accidents upward to compensate for the incompleteness of the frame from which they were selected.

The other checking method involves going to towing agencies in each PSU, recording basic information from their records on each accident vehicle which they have towed into their agency, and then checking for non-duplicates against the accident frame file. This would be a fairly time-consuming and tedious task, and in most PSUs it would probably be done only once near the beginning of the project in order to determine if there is a serious problem of towaway accidents not being available in the police agency records. If this checking process finds that fewer than 10% of the towaway accidents are not available in police agency records, then probably it would be safe to ignore the problem, especially if it is determined that these tend to involve less serious accidents and injuries. If the difference is greater than 10%, it may be necessary to establish special procedures for regularly gathering accident information from local towing agencies into a towing agency frame. The two frames would then have to be compared for duplicates, and a selection process similar to that used with the police agency accident frame would have to be established for selecting accidents to investigate from the non-duplicates in the towing agency accident frame. If the problem of missing towaway accidents in the police files were found to be quite substantial, then it might prove necessary to move to a substitute PSU, although such a step should be avoided if at all possible.

It should be mentioned that the computerizing of the police agency accident frame should be of more value than just to permit checking on the accident sample selection process. These frame data from all the PSUs will provide a large national sample of basic information about accidents of the designated types, and thus this frame sample can be used to provide more reliable statistical estimates for certain basic variables available in the police agency accident records than will be available from the smaller detailed investigation sample. There will be 100,000 to 250,000 accident cases annually in the national accident frames file, depending on the number of PSUs established, and this file could provide some useful basic information of the type which is regularly included in police accident reports. The PSU accident frames will also be useful in the selection of accidents for investigation in the Quick Response System whenever the data relating to the selection criteria for those accidents are included in the basic information recorded on the accident frame cases.

## 4.3 Special Accident Selection Procedures in Large PSUs

The accident selection process described in Section 4.2 would be expected to work most successfully in PSUs between 100,000 and 200,000 in population. In these PSUs there will be enough accidents to permit oversampling of the types of accidents of particular interest, and yet there will not be so many accidents in the police accident files that the accident investigation team will have to spend a very large portion of its time screening these files and compiling the accident selection frame. However, only about 10% of the U.S. population live in PSUs of this ideal size. About 30% live in smaller PSUs in which there may not be as many accidents to investigate as would be desired in certain subclasses, even if they are sampled at a 100% rate. Fortunately, this is not a serious problem, since even the smaller PSUs will be expected to provide significant numbers of accidents in all subclasses, and with the addition of data from the large PSUs there should be sufficient numbers in each subclass in the national sample to permit useful analysis.

The more serious problem is with the 60% of the PSUs which are expected to be larger than this ideal size, since 60% of the U.S. population lives in potential PSUs larger than 200,000 in size. These include 18% in counties larger than 1,000,000; 22% in counties between 500,000 and 1,000,000; and about 20% in counties between 200,000 and 500,000. It is obvious that in these larger PSUs the process of going through all the PSU accident records and compiling a complete frame of all eligible accidents in the PSU would be very time-consuming and expensive. Therefore it seems essential that another stage of selection be employed in these PSUs. Without more experience in following the accident selection procedures described in Section 4.2 it is difficult to judge at exactly what size PSU it would be best to make use of a secondary selection procedure. This will be one of the questions to be explored in the pilot phase of this project. Statistically it would be best if all eligible accidents within a PSU were included in the PSU frame, but in Wayne County (Detroit) this would involve sifting through more than 100,000 accident records annually and recording data for the accident frame on perhaps 40,000 of them. The small statistical gains from not employing secondary selection would hardly justify such an expensive procedure in such a large PSU.

Three major approaches to the secondary selection stage have been considered for these large PSUs. The first involves a subsampling of geographic areas within the county in relation to population size. This is the standard method followed in constructing household interview samples. A large city would be divided into several geographic sections, hopefully in relation to police precinct accident files (if there are such) or at least in relation to some kind of geographic indicator which is recorded on the accident record and which could be quickly spotted by a team member going through the central accident files. A population size would be associated with each of these geographic sections, and a controlled selection would be made among these city sections and among the other police agencies in the county, using probability-proportionate-to-size techniques. To ensure diversity in the types of areas chosen to represent the PSU, the potential Secondary Selection Units (SSUs) might be stratified according to population size, or proximity to the downtown area of the central city, or some other variable of concern, and a certain number of selections would be made from each stratum in relation to the ratio of the stratum population to the whole PSU population. Another method for providing diversity among SSUs, which would provide more compact secondary selection areas and thus logistical savings for the accident investigation team, would be to divide the PSU into several diverse geographic areas in rough pie shapes, each area to include parts of the central city as well as near-suburbs and far-suburbs. The SSUs selected in a large PSU should probably total between 200,000 and 500,000 in population. Once these SSU police agencies are chosen the same procedure described in Section 4.2 will be used to compile all eligible accidents into the PSU accident frame and to select certain accidents for the accident investigation sample.

A second secondary selection approach would involve not a permanent subsampling of police agencies and city precincts but instead would involve rotating among all the different police agencies in the PSU on different days of the year. For example, if it was decided that only one-fourth of a PSU's eligible accidents should be included in the accident frame, then each PSU police agency would be randomly assigned a particular 91 or 92 days a year, and only accidents which took place on these particular dates would be screened for inclusion in the accident frame from that agency. In determining the dates for each agency it would probably be desirable to introduce a stratification procedure for both months and days of the week. This second approach would reduce the number of accidents which would have to be compiled into the accident frame, and it would reduce the frequency of visits to each of the police agencies in the PSU. However, it still would not avoid the necessity of sifting through all of the accident records in each police agency, unless the agency happens to adhere rigidly to a procedure of filing accidents by date of occurrence. If this were not the agency procedure, all of the accident records would have to be screened for the possibility of a late-filed report covering an accident on a designated date but not filed with the other accident records for that date. Still this procedure of screening all accident records in relation to date of occurrence would be much less time-consuming than would be reading all records to see if an accident is an eligible type and then entering all eligible accidents into the accident frame.

The third secondary selection approach is similar to the second. Rather than using dates, some other variable would be used for randomly selecting among all the accident records in every policy agency in the PSU. If the police agency assigns an accident report filing number as the reports are filed, this would be a most logical basis for choosing a certain fraction of the accident reports (e.g., using all records whose filing numbers end in 1, 7, or 9 for a 30% sample). A prescribed procedure utilizing a number on the form would be preferable to having a team member count record forms and select certain ones at a designated sampling rate (every third, or fifth, etc.). In a large city, if the accident forms are assigned a sequential filing number and are really filed in order, it might be possible to use the tens digit or even the hundreds digit to select groups of records for screening while disregarding large sections of the accident files.

The second and third approaches have the statistical advantage of approximating a simple random sample of the accident records in all of the police agencies of the PSU. They have the disadvantage that they still involve use of the entire accident files of all the police agencies in the PSU. In Los Angeles County with perhaps 300,000 accidents recorded per year by 78 city police agencies, by the Los Angeles County Sheriff's Department, and by the various California Highway Patrol posts in the area, it is probably too much to use the accident files of all the police agencies in the county. Perhaps some combination of approaches, involving the reduction in the number of utilized police agency files by means of the first approach and then taking a random sample of these files by means of the second or third approaches, would be the best secondary selection method in the largest PSUs such as Los Angeles County and Cook County (Chicago). Of course it is not necessary that the same secondary selection process be used in each of the large PSUs, but the process used can be tailored to best suit each local situation. It is planned that different approaches to secondary selection be tried out during the pilot phase, so that more definitive recommendations can be made before the full sample design is implemented.

## 4.4 Weighting Procedures and the Calculation of Estimates and Their Sampling Errors

There are two types of weighting factors which must be used with each accident case in the analysis of the accident sample data. The first is derived from the sampling rate for the selection of the particular PSU.

The second is derived from the sampling rate used to select the particular case for investigation.

The probability of selection of each PSU is the ratio of the PSU's population to the average population of the non-self-representing strata (5,896,347 in the suggested plan of 32 non-self-representing PSUs). The inverse of this probability of selection of each PSU should be used to weight each data case from that PSU when it is entered into the national sample. Thus in the suggested plan the population weighting factor for cases in a PSU of 1,000,000 people would be 5.896 (the inverse of the 1,000,000/5,896,347 probability of selection), and the population weighting factor for cases in a PSU of 100,000 would be 58.96 (the inverse of the 100,000/5,896,347 probability of selection for that PSU). In large PSUs using secondary selection procedures, this population weighting factor would be increased by multiplying by the inverse of the secondary selection sampling weight. For example, if only half of the accidents were selected for screening in a PSU of 1,000,000 population, then the population weighting factor would be 2 x 5.896 = 11.79. Population weights in self-representing PSUs would similarly be based on the ratio between the average population of the non-self-representing strata and the PSU population.

The second weighting factor would be the inverse of the particular sampling fraction used to select an eligible accident for investigation as representative of a particular subclass of accidents. For example, if, in a particular PSU, 50% of the eligible late-model, hospitalized-victim accidents were chosen for the accident sample, then each such investigated case would be weighted by a factor of two. If 20% of the pedestrian accidents were included in the accident sample, each such case would be weighted by a factor of five.

Thus the population weighting factor would be identical for every case in the accident sample from a given PSU, while the case selection weighting factor would be identical for each case in a given subclass of sampled accidents in a given PSU. Both weighting factors would be known at the time of the case selection and would be included as part of the case data record from the beginning. For use in analysis programs the two factors would be multiplied to form a single sampling weight variable, and to provide more manageable weighted frequencies it would probably be desirable to divide this variable by some large constant (say 100 or 200). All of the relevant analysis programs can make use of decimal weights just as well as integer weights. The population weighting factor should be used in the analysis of data in the national accident frame as well as in the analysis of the national accident sample.

Both the national accident frame and the national accident sample will be analyzed by computer programs to produce various national and regional estimates of total frequencies, means, proportions, ratios, etc. For calculating the variance and sampling error of these various estimates, The University of Michigan Institute for Social Research has a set of computer programs developed by Leslie Kish, Martin Frankel, and Neal Van Eck.* These programs were written to calculate variances, sampling errors, and design effects for national samples of the type suggested in this report. The most useful

*SEPP: Sampling Error Program Package, Ann Arbor: Institute for Social Research, no date (about 1971).

program in the package is PSALMS (Paired Selection Algorithm for Multiple Subclasses). It computes sampling errors using a method based on the Taylor approximation for simple ratios and for linear combinations of ratios (for example, differences of means or proportions between two subclasses). The BRRP program (Balanced Repeated Replication Package) uses the method of balanced repeated replications, and it can also compute sampling errors for various types of correlation and regression coefficients. Both programs are prepared to receive data containing a weight variable, although there will always be some question about the interpretation of sampling errors of correlation and regression coefficients when using weighted data.

Both programs are based on a model utilizing comparisons between and among pairs of similar PSUs, although these can either be a distinct set of pairs or an ordered set of PSUs in which the first is paired with the second, the second is paired with the third, etc. The program also permits the input data to come from a combination of these two types of pairings. The introductory text for the SEPP programs and the sampling error formulas used by PSALMS are included in Appendix C.

It would be desirable in the projection of national estimates of total frequencies from the accident sample or frame to be able to compare some of the results with known parameters. One example of such an independently available national parameter which has high reliability is the number of fatal accidents in the nation. The NASS estimate of the total number of fatal accidents for a certain period could be compared with the known national number of fatal accidents (excluding Alaska and Hawaii) in order to judge the degree of representativeness of the data in the national accident frame and sample. If there were a significant difference among the two figures, the ratio of this difference could be used for adjusting all of the national estimates of total frequencies which are based on the national accident frame or sample.

FIGURE 4.1 ESTIMATED RELATIVE SAMPLING ERRORS FOR DIFFERENT NUMBERS OF PSUS AT A $6,000,000 COST ESTIMATE, 8 LOWER ERROR VARIABLES IN 1973 TEXAS 5% SAMPLE DATA

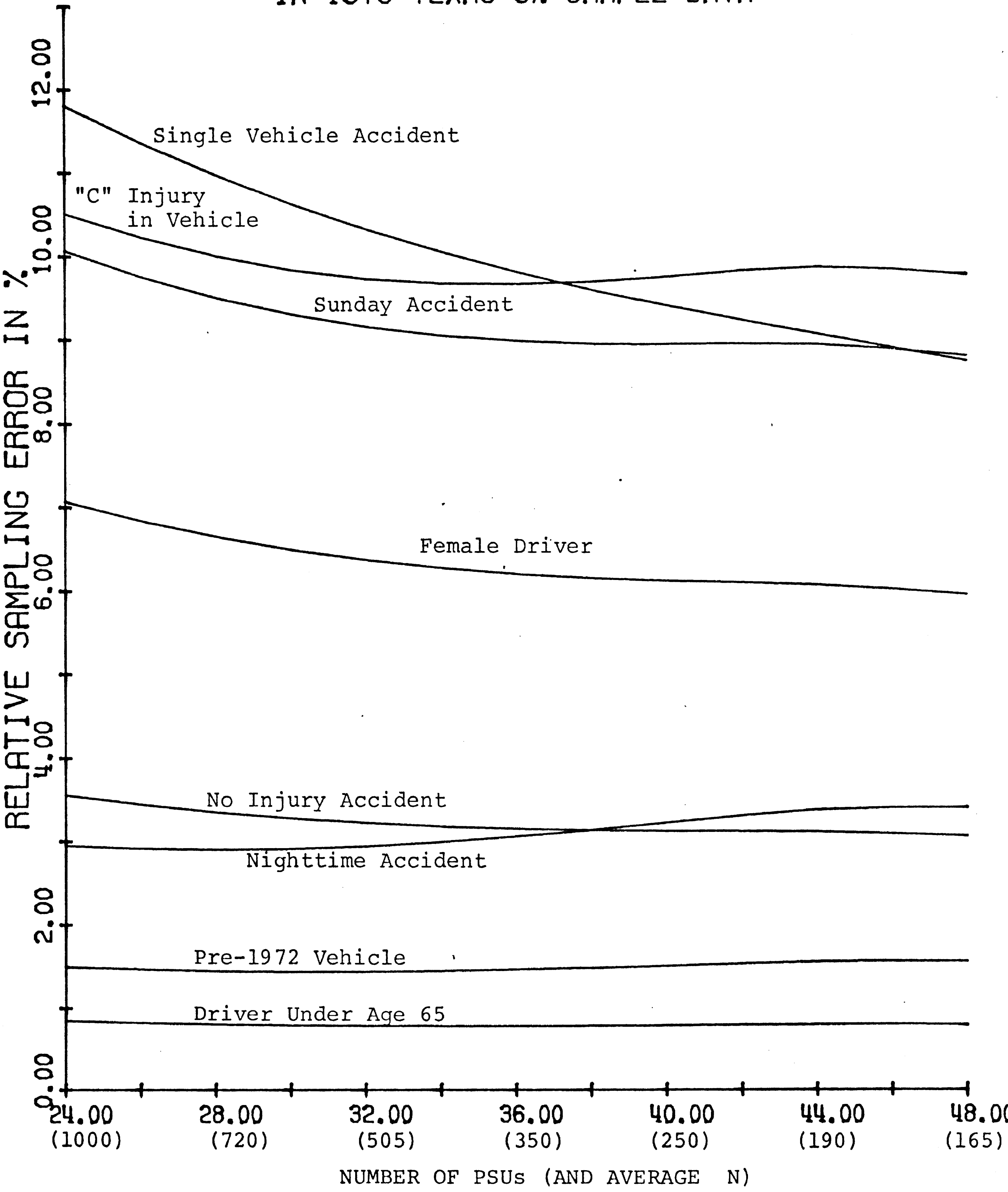


FIGURE 4.2 ESTIMATED RELATIVE SAMPLING ERRORS FOR DIFFERENT NUMBERS OF PSUS AT A $6,000,000 LOST ESTIMATE, 5 HIGHER ERROR VARIABLES IN 1973 TEXAS 5% SAMPLE DATA

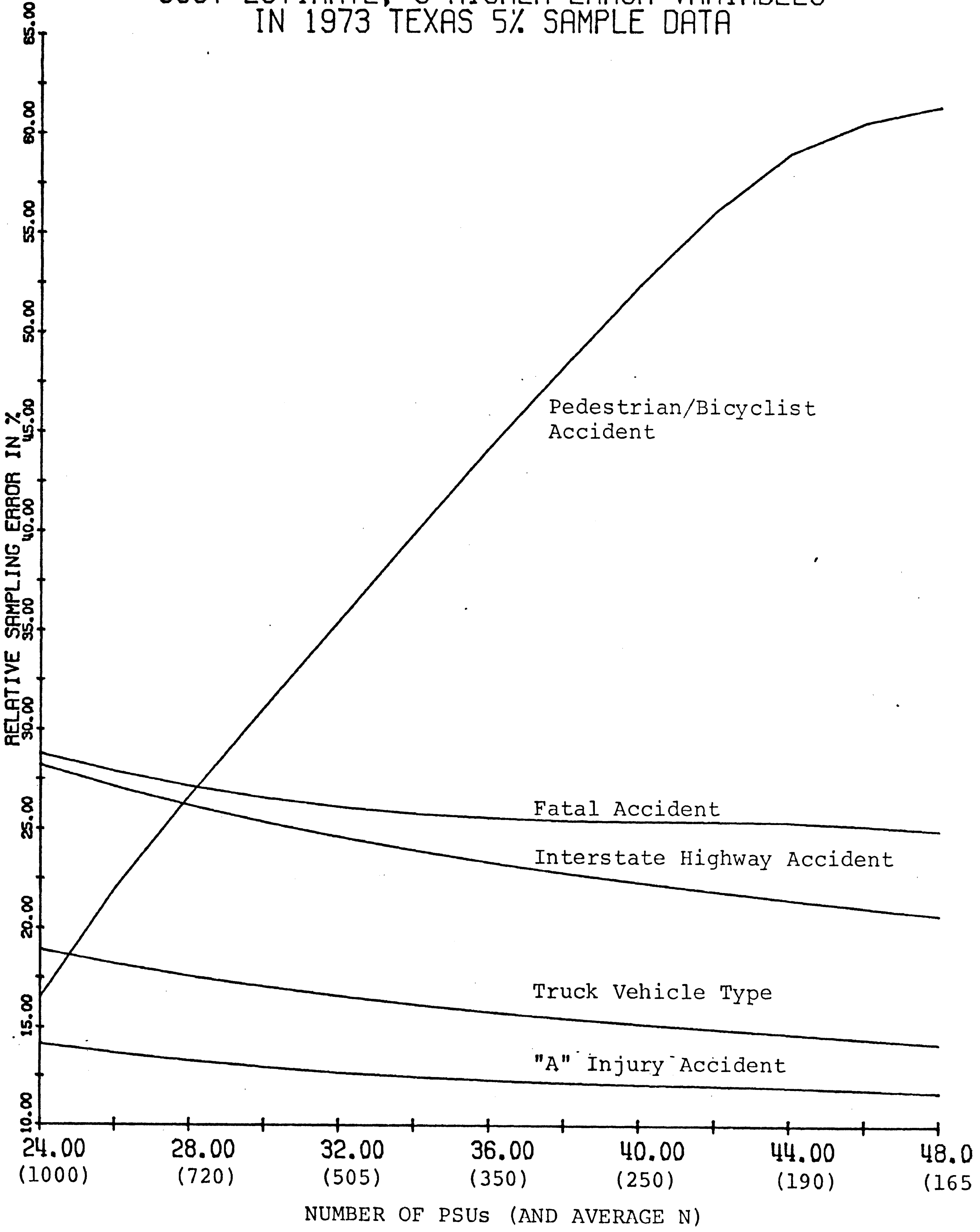


FIGURE 4.3 ESTIMATED RELATIVE SAMPLING ERRORS FOR DIFFERENT NUMBERS OF PSUS AT A $6,000,000 COST ESTIMATE, 8 LOWER ERROR VARIABLES IN 1973 MICHIGAN 5% SAMPLE DATA

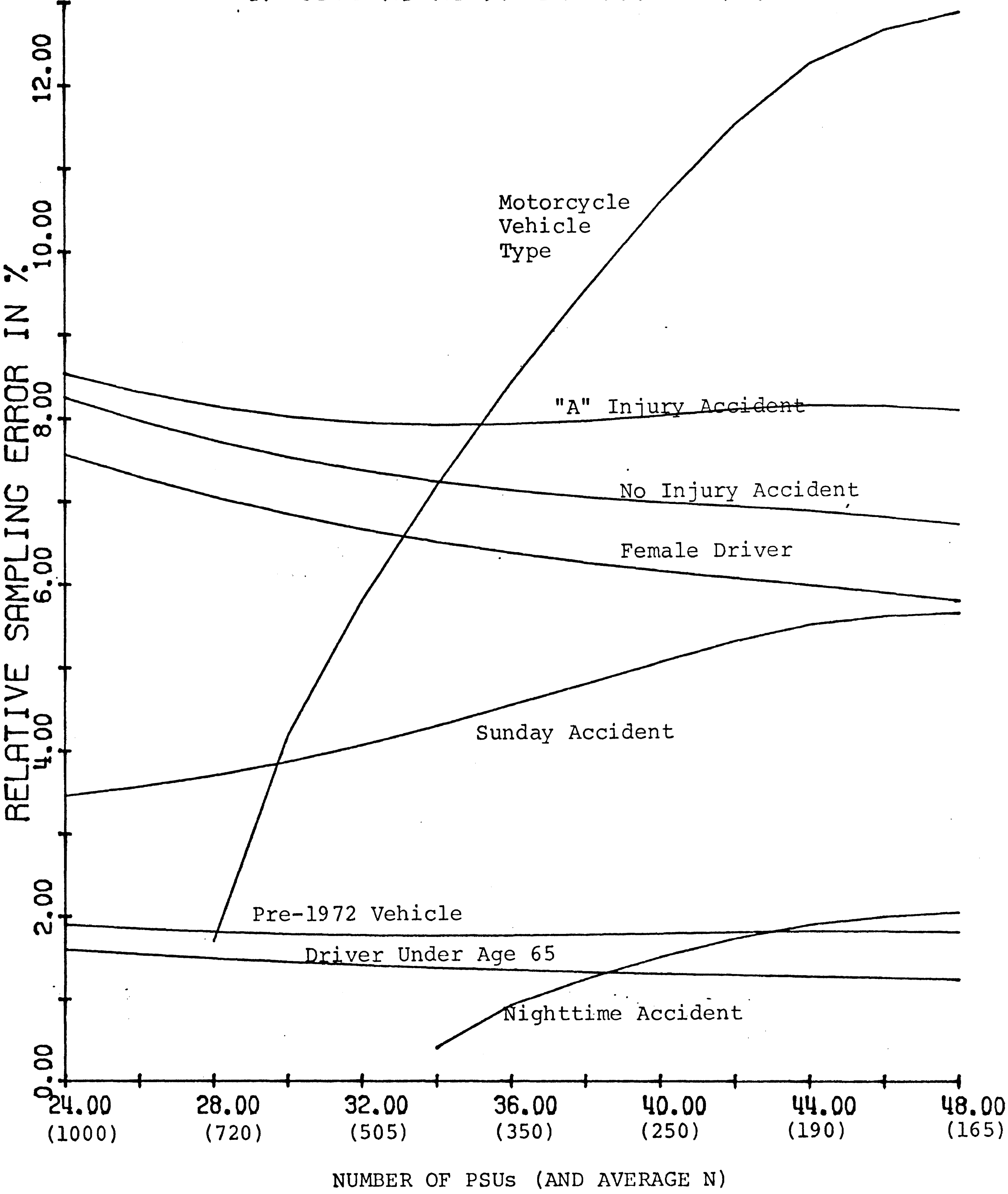


FIGURE 4.4 ESTIMATED RELATIVE SAMPLING ERRORS FOR DIFFERENT NUMBERS OF PSUS AT A $6,000,000 COST ESTIMATE, 6 HIGHER ERROR VARIABLES IN 1973 MICHIGAN 5% SAMPLE DATA

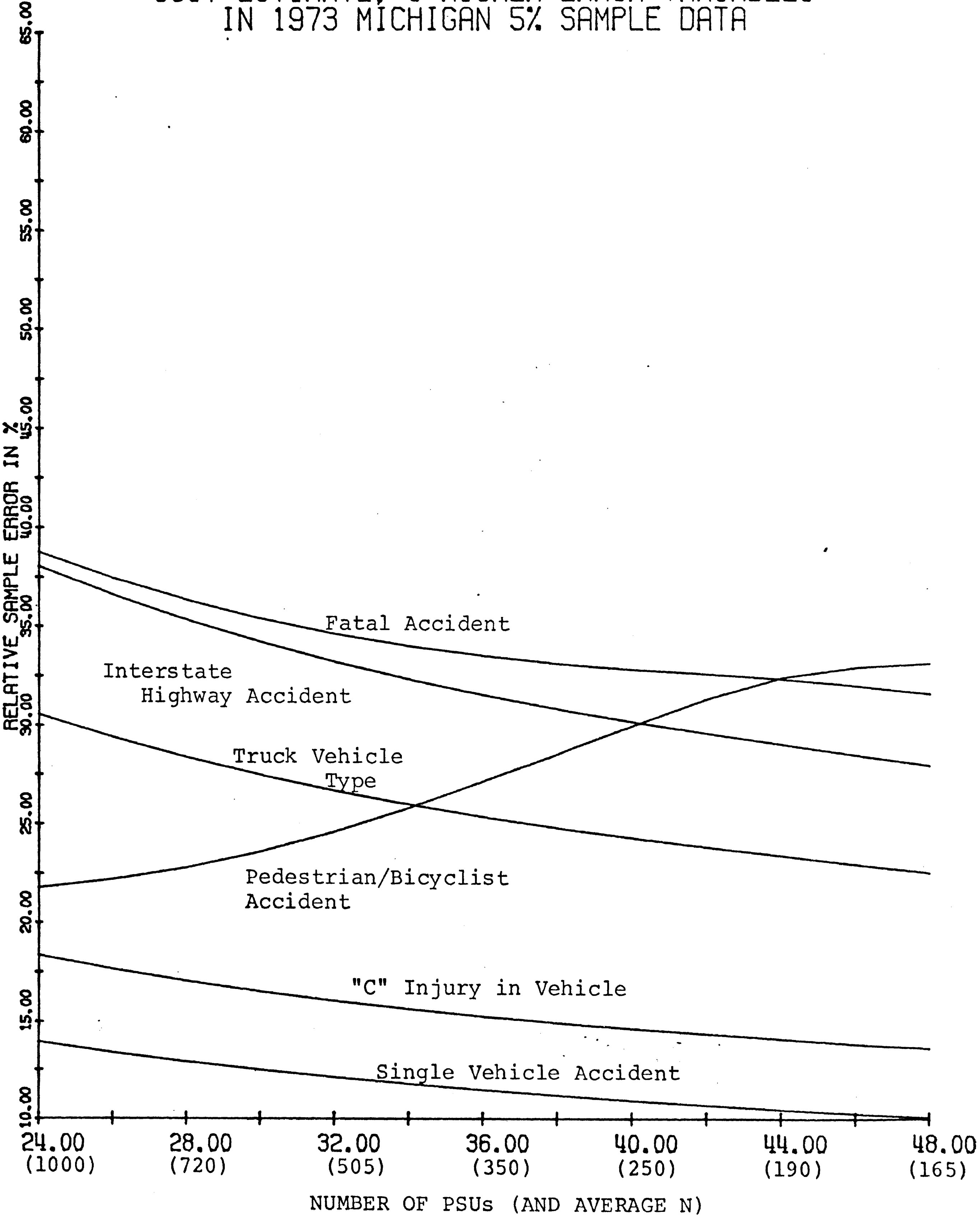


FIGURE 4.5 ESTIMATED RELATIVE SAMPLING ERRORS FOR DIFFERENT NUMBERS OF PSUS AT A $12,000,000 COST ESTIMATE, 9 LOWER ERROR VARIABLES IN 1973 TEXAS 5% SAMPLE DATA

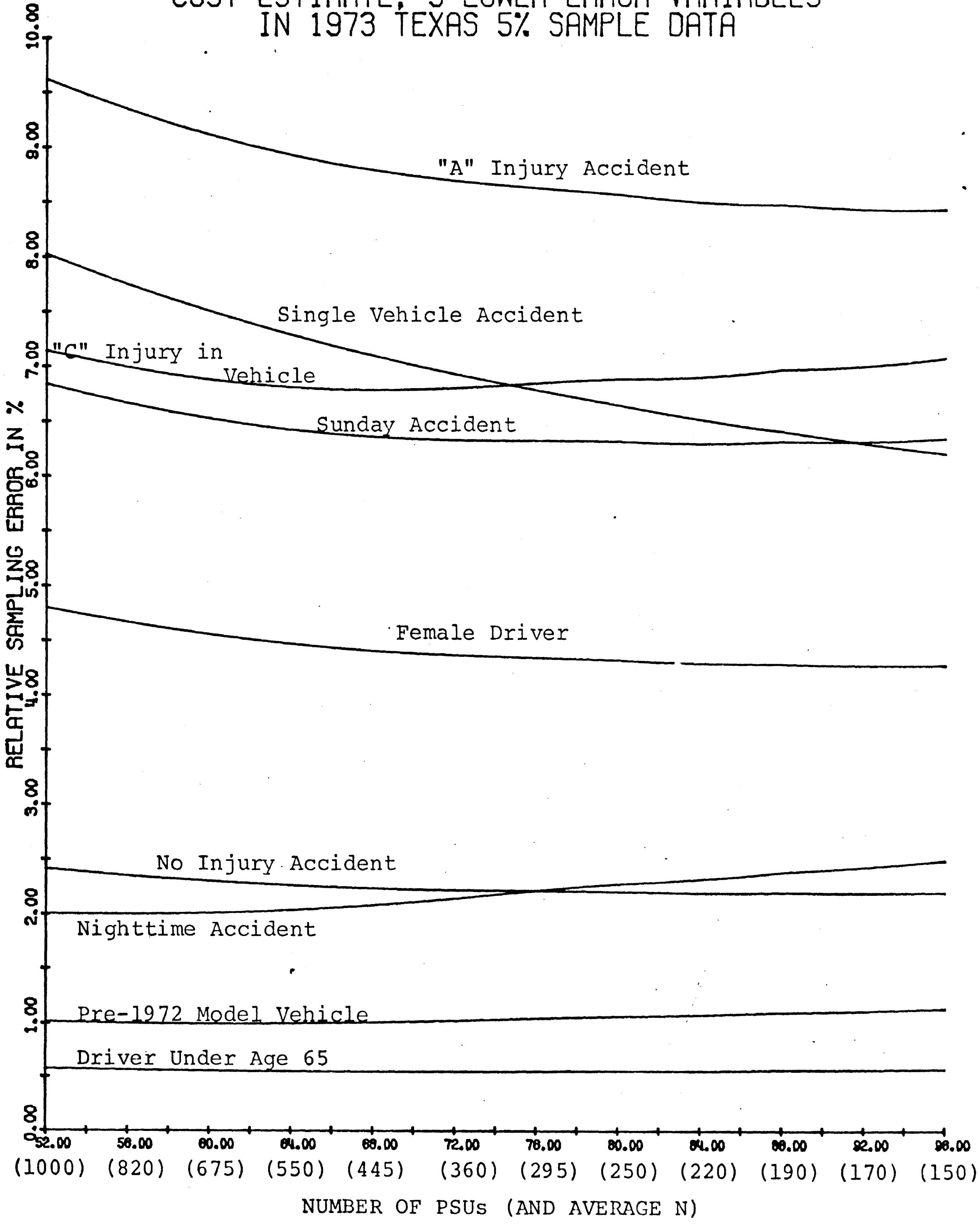


FIGURE 4.6 ESTIMATED RELATIVE SAMPLING ERRORS FOR DIFFERENT NUMBERS OF PSUS AT A $12,000,000 COST ESTIMATE, 4 HIGHER ERROR VARIABLES IN 1973 TEXAS 5% SAMPLE DATA

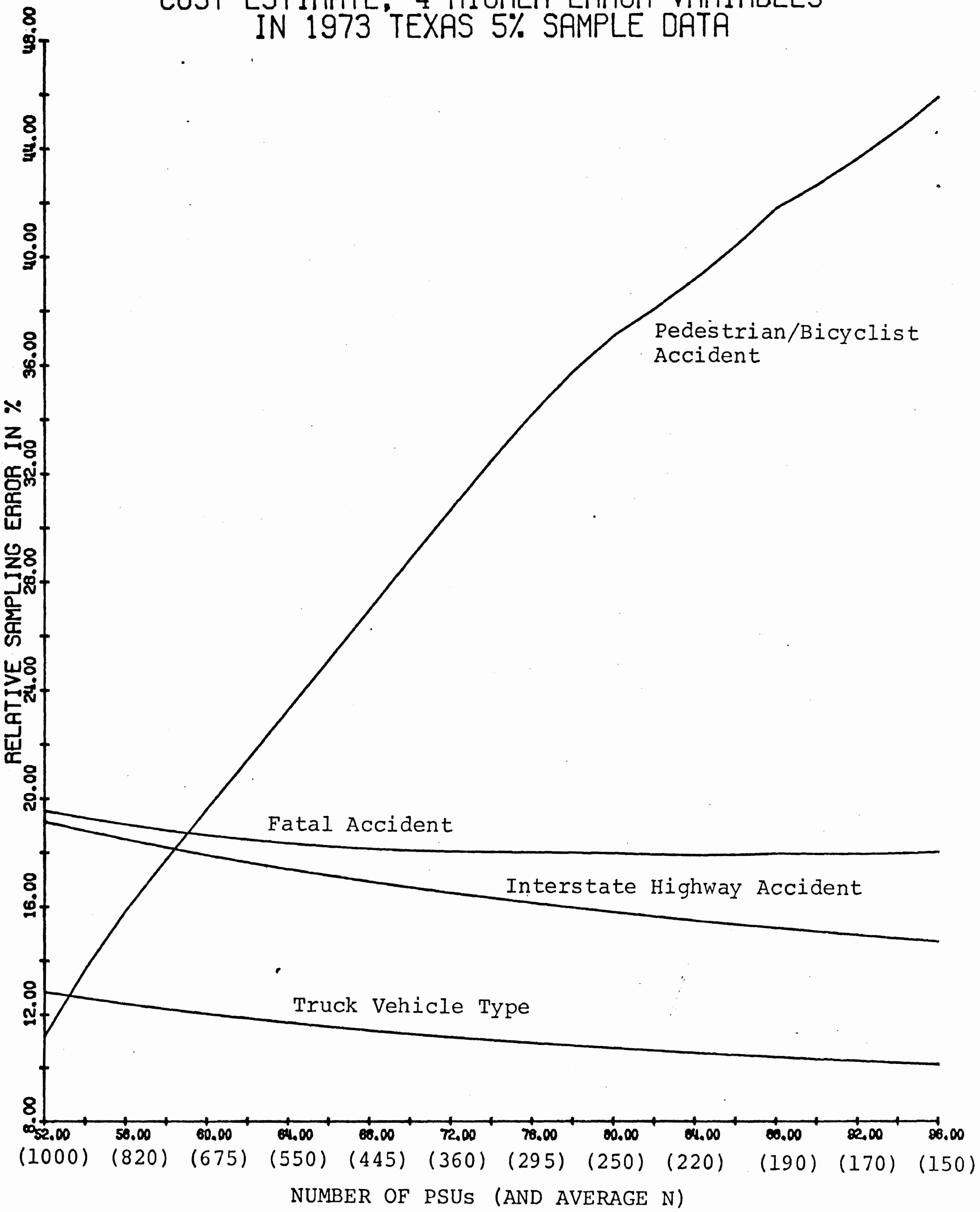


FIGURE 4.7 ESTIMATED RELATIVE SAMPLING ERRORS FOR DIFFERENT NUMBERS OF PSUS AT A $12,000,000 COST ESTIMATE, 9 LOWER ERROR VARIABLES IN 1973 MICHIGAN 5% SAMPLE DATA

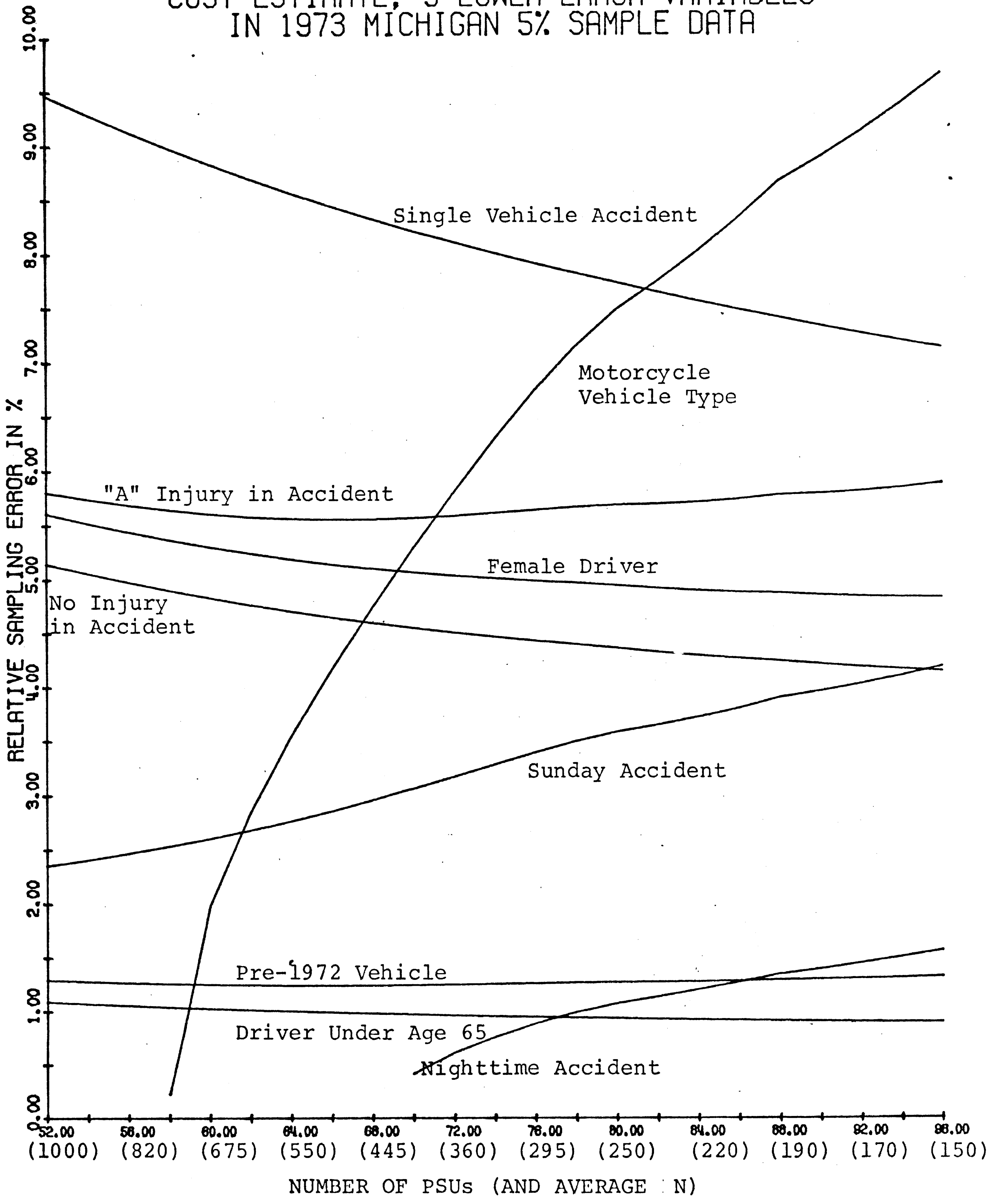
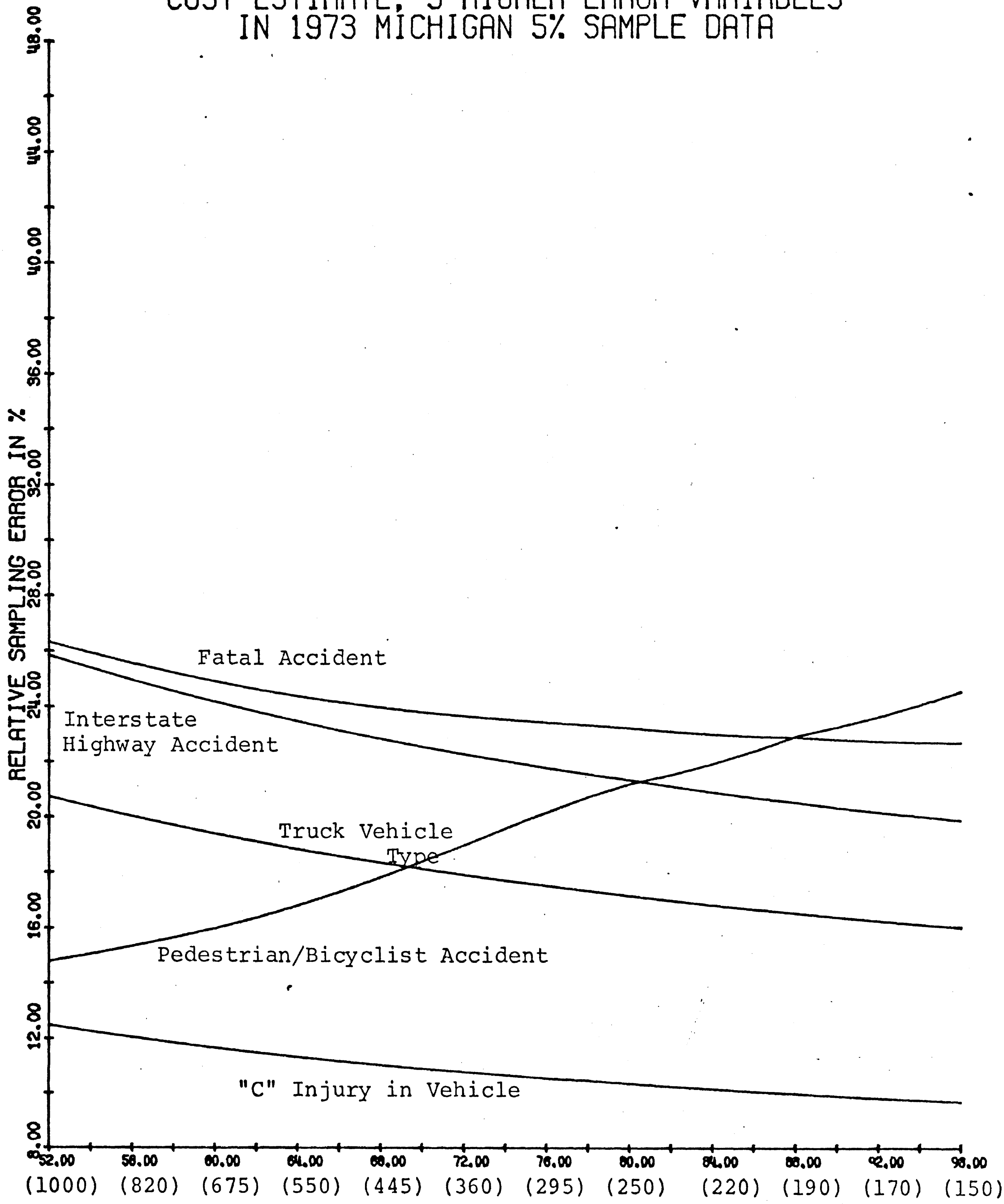


FIGURE 4.8 ESTIMATED RELATIVE SAMPLING ERRORS FOR DIFFERENT NUMBERS OF PSUS AT A $12,000,000 COST ESTIMATE, 5 HIGHER ERROR VARIABLES IN 1973 MICHIGAN 5% SAMPLE DATA



# 5. CONSTRUCTION OF THE NATIONAL SAMPLE

To construct a national sample of PSUs in accordance with the design discussed in Section 4.1, it was necessary to obtain three types of data about each county in the 48 contiguous states. These were its population, its retail sales at gasoline service stations, and whether or not it was part of an SMSA. The source used for the population data was General Revenue Sharing: Initial Data Elements, Entitlement Period Six, published by the Office of Revenue Sharing of the Department of the Treasury in April, 1975. This volume contains population estimates as of July 1, 1973, for every local governmental unit in the United States--counties, cities, villages, towns, and townships. The Census Bureau did not complete its publication of provisional 1974 estimates of county populations in time for their utilization in the sample construction. The source of data on gasoline sales was Table 6 in the series of state reports on the 1972 Census of Retail Trade, issued by the Census Bureau in 1973. The source of information concerning SMSA composition was Federal Information Processing Standards Publication 8-4, issued June 30, 1974. This contained information defining the 267 SMSAs which contain almost three quarters of the U.S. population. It should be noted that there has been a considerable expansion in the number of SMSAs and in the geographic boundaries of many SMSAs since the 1970 Census.

## 5.1 Creation of the Potential PSUs

As discussed earlier, the sample design calls for the selection of PSUs no smaller than 50,000 in population. Therefore a laborious process had to be used to group smaller counties together into minimum-sized potential PSUs. This involved xeroxing outline maps of each state; marking off the counties in each SMSA in each state; entering the 1970 population figure to the nearest thousand in each county on each map; and then trying to put the smaller counties together in as compact and natural groups as possible. To try to make these groupings natural ones based on normal distribution centers, state road maps and the Rand McNally map of basic trading areas were consulted. However, it was often difficult to use natural trading areas, because most small distribution centers tend to serve mainly the county in which they are located, while larger distribution centers tend to be located in counties large enough to serve as a potential PSU without grouping.

In grouping the PSUs the four regional divisions were adhered to completely, and there was only one potential PSU which was constructed across state lines. Three small (in population) California counties located on the eastern slopes of the Sierra Mountains were joined with seven counties in western Nevada, because they are much more accessible from Nevada during much of the year. The main rationale for not crossing state boundaries in the PSU groupings was to avoid having to involve more than one state police force or highway patrol as a source of PSU accident records.

However, for the rural-urban stratification factor it was not possible to follow the boundaries of the three rural-urban strata so strictly. In 36 of the SMSAs there were SMSA suburban counties too small to serve as PSUs by themselves and yet there were no other nearby suburban SMSA counties to group them with. So these were grouped with their central county, and these groups were all classified in the SMSA central county stratum, making this sttatum slightly larger in total population than it should be. There were also two cases in which geographic logic necessitated the combination of some small non-SMSA counties with a central SMSA county. There were also 14 cases in which non-SMSA counties were combined with SMSA suburban counties. These were primarily situations in which a small suburban SMSA county was located across a state boundary from the main parts of the SMSA and there were no other nearby suburban SMSA counties to combine it with. These county groups were assigned to whichever stratum was appropriate for the majority of the group population.

In determining appropriate rural-urban strata assignments, the procedure used with double-named or triple-named SMSA's should be mentioned. If named central cities were in different counties and each central city was larger than 50,000 in population, then each county was considered an SMSA central county. However, if the named city of a second county was less than 50,000 in population, that county was considered a suburban SMSA county. For example, both Mahoning and Trumbull Counties were considered central counties in the Youngstown-Warren SMSA (Ohio); but Greenville County was considered a central county while Spartanburg County was considered a suburban county in the Greenville-Spartanburg SMSA (South Carolina).

Determining central county and suburban county assignments in the New England states was especially complicated, because in these states SMSAs are defined by cities and towns rather than by whole counties. Only in four large SMSAs were separate central county and suburban potential PSUs established: Boston, Hartford, Springfield, and Providence. In these SMSAs, large parts of the SMSA in different counties from the central city were considered as suburban SMSA counties. All other SMSAs were treated as single central county potential PSUs.

After the potential PSUs were defined geographically, the components of each potential PSU were listed on separate forms. Since neither the 1973 populations nor the 1972 retail trade data were yet available on computer tapes from the Census Bureau, it was necessary to look up these data in the documents mentioned earlier, to record the two figures for each component of a potential PSU, and then to sum the two sets of figures on a calculator. There were no special difficulties with this clerical process for the population data. However, for a number of small counties the gas sales data were not published for reasons of confidentiality, although the numbers of gasoline service stations in these counties were published. Therefore an estimate of gasoline sales for such a county was made by multiplying the state average gas sales per service station by the number of service stations in that county. This problem was even greater in New England, where data were not published for a large number of the smaller cities and towns. Here the basic procedure used was to prorate the published figure for gasoline sales in the "remainder of the county" among the remaining county units in proportion to their populations.

After the input data were totaled for each potential PSU, the forms were keypunched onto one IBM card per PSU. This card contains a PSU identification number, which includes both a region and a rural-urban code, the number of counties in the PSU, and the total 1972 retail gasoline sales in the PSU. After being keypunched the data records were entered into a computer file, and each of the variables was summed by state in order to check them against the published state totals. This process resulted in discovering several small clerical errors, which were then corrected.

## 5.2 Selection of the National Sample of PSUs

The first step in using the data records for the 1,241 potential PSUs in order to select a national sample involved the creation of a new variable--the per capita gas sales--by dividing the total gas sales by the population. Percentile distributions were then obtained on this variable in order to decide what strata categories to establish. Since most of the areas which are high in per capita gas sales tend to be rather low in population, it was decided to use the 20th, 40th, and 60th percentiles as cut-off points in establishing four gas sales categories (low, moderate, average, and high). The three cut-off values for these four categories are $138, $158, and $178 per capita gas station sales. The full range of values runs from $62 per capita in New York City to $553 per capita in seven Nevada counties.

With this third stratification variable created, it was then possible to group the potential PSUs into their appropriate controlled-selection cells and to have the computer calculate the population totals for each cell. Before beginning the controlled-selection process it was decided that the three largest potential PSUs, New York City, Los Angeles County, and Cook County, should be selected with certainty, as self-representing PSUs. These populations were then taken out of their assigned strata, leaving 1,238 potential PSUs with a total population of 188,683,119 allocated among 48 controlled-selection cells. It was also decided that in the recommended selection process presented here there would be 32 non-self-representing PSU selections, with exactly eight from each of the four regions. An adjustment was made in the boundary between the Northeast region and the South region by moving Delaware, Maryland, and Washington into the Northeast. This change left 25.06% of the non-self-representing PSU population in the Northeast, 24.87% in the Midwest, 24.51% in the South, and 25.56% in the West (defined as the Pacific States, the Mountain States, and the six Great Plains States, including Texas). Thus, constraining the selection probabilities for each stratum to permit exactly eight selections from each region caused little deviation from the non-constrained probabilities. The major reason for this constraint was to ensure that PSU pairings for the calculation of sampling errors could be carried out separately in each region.

Table 5.1 presents the constrained selection probabilities for each of the 48 strata defined by the intersection of the four regional strata, the three urban-rural strata, and the four gasoline sales strata. The next step in the selection process involved entering these probabilities into a special controlled-selection computer program recently developed by Robert Groves and Irene Hess of the Institute for Social Research Sampling Section. This program ascertains all of the possible selection allocation patterns among the 48 strata which are appropriate within the constraints of the given marginals for each of the four regions. It also assigns a probability weight to each pattern and cumulates these weights as each new pattern is presented. Thus a random number table can be used to pick a number between one

TABLE 5.1. SELECTION PROBABILITIES AND THE NUMBER OF ACTUAL SELECTIONS IN THE CHOICE OF 32 NON-SELF-REPRESENTING PSUs FROM 48 CELLS

| | Gasoline Sales Strata | | | | |
|---|---|---|---|---|---|
| Northeast | 1 | 2 | 3 | 4 | T |
| Urban - 1 | 2.0265 (2) | .8316 (1) | .7310 (1) | .4239 (0) | 4.0130 (4) |
| Rural - 2 | .5098 (1) | .7801 (1) | .4327 (0) | .8145 (0) | 2.5371 (2) |
| Strata - 3 | .2882 (0) | .3712 (0) | .3695 (1) | .4210 (1) | 1.4499 (2) |
| T | 2.8245 (3) | 1.9829 (2) | 1.5332 (2) | 1.6594 (1) | 8.0000 (8) |
| Midwest | 1 | 2 | 3 | 4 | T |
| Urban - 1 | .0206 (0) | 1.4213 (1) | 1.2940 (1) | 1.0352 (1) | 3.7711 (3) |
| Rural - 2 | .1496 (0) | .2784 (0) | .4750 (1) | .8407 (1) | 1.7437 (2) |
| Strata - 3 | .1882 (0) | .3090 (1) | .5406 (0) | 1.4474 (2) | 2.4852 (3) |
| T | .3584 (0) | 2.0087 (2) | 2.3096 (2) | 3.3233 (4) | 8.0000 (8) |
| South | 1 | 2 | 3 | 4 | T |
| Urban - 1 | .5312 (0) | .7850 (1) | 1.1505 (2) | 1.0127 (1) | 3.4844 (4) |
| Rural - 2 | .3018 (0) | .1957 (0) | .1917 (0) | .4254 (1) | 1.1146 (1) |
| Strata - 3 | 1.3510 (2) | .8743 (1) | .4265 (0) | .7492 (0) | 3.4010 (3) |
| T | 2.1840 (2) | 1.8550 (2) | 1.7687 (2) | 2.1923 (2) | 8.0000 (8) |
| West | 1 | 2 | 3 | 4 | T |
| Urban - 1 | .4686 (1) | 1.0738 (1) | 2.2313 (2) | 1.0110 (1) | 4.7847 (5) |
| Rural - 2 | .1263 (0) | .3528 (0) | .2846 (0) | .0765 (1) | .8402 (1) |
| Strata - 3 | .1839 (0) | .1770 (0) | .2560 (1) | 1.7582 (1) | 2.3751 (2) |
| T | .7788 (1) | 1.6036 (1) | 2.7719 (3) | 2.8457 (3) | 8.0000 (8) |
| Nation | 1 | 2 | 3 | 4 | T |
| Urban - 1 | 3.0469 (3) | 4.1117 (4) | 5.4068 (6) | 3.4878 (3) | 16.0532 (16) |
| Rural - 2 | 1.0875 (1) | 1.6070 (1) | 1.3840 (1) | 2.1571 (3) | 6.2356 (6) |
| Strata - 3 | 2.0113 (2) | 1.7315 (2) | 1.5926 (2) | 4.3758 (4) | 9.7112 (10) |
| T | 6.1457 (6) | 7.4502 (7) | 8.3834 (9) | 10.0207 (10) | 32.0000 (32) |

and the cumulative weight total for the final pattern, and this number then determines which selection pattern is to be utilized. The controlled selection process is described more fully in an article by Roe Goodman and Leslie Kish entitled, "Controlled Selection -- A Technique in Probability Sampling," Journal of the American Statistical Association, 45:350-372 (September, 1950), and in Section 12.8 of Survey Sampling by Leslie Kish (New York: Wiley, 1965, pp.488-495).

When the probability data in Table 5.1 were input to the Groves-Hess controlled-selection program, 38 possible selection patterns were generated. By random number selection the fifteenth pattern happened to be chosen, and the selection allocations of this pattern are also shown in Table 5.1.

It is appropriate to choose similar pairs of the selected cells for use in the sampling error programs prior to the actual selection of particular PSUs. The 16 suggested pairs for the selection pattern of Table 5.1 are shown in Table 5.2. In the sampling error calculations the PSUs chosen to represent these cells would be paired, and the three self-representing PSUs would be treated as consecutive pairs. A random process of choosing one PSU from each of these pairs would be a good method for obtaining a representative national sample of 16 PSUs to use in the second year of implementing the sample plan.

If it is decided that either 35 or 67 PSUs are appropriate numbers to be established in the national Continuous Sampling Subsystem, then the selection allocation shown in Table 5.1 shows which cells should have one or two selections. The potential PSUs in each cell are listed in Appendix D in order by number of counties and decreasing population size. This listing

TABLE 5.2. SUGGESTED PAIRINGS OF CELL PSU SELECTIONS FOR THE CALCULATION OF SAMPLING ERRORS, USING THE SELECTION PATTERN OF TABLE 5.1.

Northeast

Pair 1 - 2 PSUs from Cell 11
Pair 2 - Cells 12 and 13
Pair 3 - Cells 21 and 22
Pair 4 - Cells 33 and 34

Midwest

Pair 1 - Cells 12 and 13
Pair 2 - Cells 14 and 24
Pair 3 - Cells 32 and 23
Pair 4 - 2 PSUs from Cell 34

South

Pair 1 - Cells 12 and 14
Pair 2 - 2 PSUs from Cell 13
Pair 3 - 2 PSUs from Cell 31
Pair 4 - Cells 32 and 24

West

Pair 1 - Cells 11 and 12
Pair 2 - 2 PSUs from Cell 13
Pair 3 - Cells 14 and 24
Pair 4 - Cells 33 and 34

includes the cumulative population totals for each potential PSU within it's cell, so the actual selection could be quickly carried out with a table of 8-digit random numbers by selecting for each cell with one selection a random number between one and the cumulative total for the cell and then seeing which potential PSU that number falls into. In the five two-selection cells, it would be desirable to first divide the cells into two equal half-cells and then to choose separate random numbers for each half-cell.

In the event that 64 non-self-representing PSUs are to be selected, the same procedure could be followed by using a paired selection technique in which a number which is half of the total cell population is added to or subtracted from the selected random number and then used to make a second selection within the cell. This second selection in the cell could also be used as a random substitute for the first selection in the system of 32 non-self-representing PSUs, if for some reason the first selected PSU proved unsuitable. However, it is hoped that no such substitutions would be necessary.

If a national sample containing a different number of PSUs were desired, then the new cell selection probabilities for that number of selections would have to be entered into the controlled-selection program, and a new pattern of selection allocations would be chosen.

It is expected that the designated system of PSUs will be maintained for several years without change. The only value in rotating PSU selections in the short run would be to reduce the variance in data aggregated over several years, but this possible advantage seems to be heavily outweighed by the disadvantage that comparison of year-to-year trends would be less meaningful if there were frequent changes in the PSUs from which the accident data are collected. And of course there are great operational advantages to keeping field investigation teams in their established localities for substantial periods of time.

Nevertheless it is recognized that at some time (perhaps following the 1980 Census) there will be enough changes in population distributions that some changes in PSUs will be needed. It is anticipated that such changes will involve as few movements of accident investigation teams as possible, and that such movements would take place in gradual phases. A technique for adjusting to changes in population distribution while minimizing the changes in selected PSUs is described in Section 12.7 of Survey Sampling by Leslie Kish (New York: Wiley, 1965, pp.483-488).

## 6. DATA SYSTEM DESIGN

Four kinds of data are to be collected under the NASS program. These are:

(1) CSS - Continuous Sampling System
Standard set of information.

(2) QRS - Quick Reaction System
Special sets of information for specific requirements.

(3) Sampling Frame for CSS and QRS
Set of information for sample selection.

(4) MDAI - Multidisciplinary Accident Investigations
In-depth set of information.

Too frequently data system designs begin with the design and implementation of a field data form. Only after field data collection has started is computerization of the data considered. After the fact, an analyst is asked to interpret the collected data. The problems with that approach are obvious.

The reverse approach was used in designing the NASS data system. The end products or analysis results desired were determined first, and then the design proceeded backwards through each stage of the data system, ultimately to the field data form. The remainder of this section discusses the considerations given to what and how data should be collected. The subsequent four sections describe the data system in terms of Field Data Collection (Section 7), Data Center Operations (Section 8), System Monitoring (Section 9), and Data Analysis and Reporting (Section 10).

## 6.1 Continuous Sampling System - CSS

The two major considerations in designing the CSS data collection system concerned what to collect and how to collect it. Each will be discussed in turn.

### 6.1.1 What Data to Collect

The process of deciding what data elements to collect began by compiling a list of principal questions currently being asked (largely by DOT staff). This compilation process was commenced early in the project because of its obvious interactions with the sample design process.

The proposed list of CSS questions was continually reviewed and updated to ensure comprehensiveness and relevancy. A general estimate of the relative importance or priority of each question was also made. The questions tabulated in Appendix E are listed in order of relative importance assigned subjectively by project personnel.

Some rare accident events may be easy and desirable to record and tabulate (e.g., injury distribution of school bus occupants), but they occur so infrequently over a period of several years that no accurate national estimates can be made within that time period. Therefore, a coarse estimate of the precision likely for each answer was made, using a simple computation of the precision with which a national estimate of each quantity could be made with the NASS sample size coupled with the time period over which that precision applies. This was estimated at a 95% confidence interval and assumed a design effect (i.e., the ratio of the total variance to the simple random sample variance) of 2.25.

The next steps in deciding what data to collect involved considering what computer data variables were required to answer each question and then what field data elements were required. Note that the set of computer variables and field elements is not identical, because the computer is able to derive necessary variables from field data. These derivatives may be as simple as deriving the "day-of-the-week" from the field-reported data, or as intricate as computing a crash severity score based upon a regression model of a dozen field-reported crash data elements. Wherever possible, field data elements were dropped where the computer could be taught to derive the required variable.

Consideration was also given to the ability to collect each data element in the field. If all of the previously discussed conditions are satisfied but the data element could not be reliably collected in the field, the element was dropped from further consideration. The ability to adequately collect a field data element was decided on the basis of four parameters: correctness, consistency, completeness, and cost. Thus, if a field element could not be measured accurately (e.g., the distance from the steering wheel hub to the rear light), or consistently between PSU teams, or without a lot of missing data, or without exorbitant cost in time (e.g., car disassembly), it was dropped. In the present design, for example, one of the initial questions concerned determination of the extent of carbon monoxide involvement in accidents. It was judged that information relevant to this factor could not be reliably obtained in the Continuous Sampling System. Consequently, the collection of any specific "degree of carbon monoxide poisoning" data has not been planned.

While a top-down design approach was used, from analysis questions to field data elements, it is clear that there are strong interactions of all levels of the data system that are part of its design. Just as a field investigator is unlikely to know what the statistician requires, the analyst cannot design a functional data system outside of the operational environment of field accident investigations. Just as the prototype data system designed here was an interactive process, an exercise of the pilot design will be required in order to arrive at a completely functional design.

### 6.1.2 How to Collect Data

Data collection for the CSS element of NASS, as the name implies, requires a continuous, uninterrupted collection of accident data in conformance with the sampling plan. The data collection system should have all of the obviously desirable features of ease of use, reliability, timeliness (i.e., quickness), responsiveness, adaptiveness to changing needs. The prototype design presented here exemplifies what can be done. Subsequent trial experience will provide the information necessary for "fine tuning" the system.

The design approach taken to the data collection was to (1) remove as many intermediate steps as practical, and (2) remove as much "computer coding" burden from the field investigator as practical. This approach resulted in the design and development of an interactive computer terminal data entry program and a modular field data form with semistructured data elements oriented towards the accident investigation protocol.

The CSS field form was designed in a modular format, so that, for example, the data collected from a vehicle inspection is recorded on the "vehicle module." Each data form is semi-structured so that the investigator can record what he sees without having to worry simultaneously about the specific encoding of computer variables. The field form is a "scratch pad" for taking notes about specific aspects of the investigation. It is not a computer coding form.

The completed form then serves as the investigator's field notes for inputting data into the computer system. This could be done from each PSU by the individual investigators or by forwarding the hard-copy forms to a central facility for encoding by a core of data editors. A coding handbook complements the field form by defining the acceptable responses to each of the data elements. Wherever possible, responses are standardized (e.g., Y, N, ?, N/A) or in natural language (e.g., "fire"). Mnemonic responses (e.g., "F" for "fire") are also accepted. The resulting field data collection protocol is described in more detail in Section 7.0.

Entry of field data into computer files is done directly from local terminals at each PSU, using computer query programs that permit the investigator to enter the data in an interactive mode. Being interactive, the data entry program permits the investigator to request descriptions of proper responses, to propose new responses where needed, to list and review cases as they are entered, and to edit or correct previous entries. On-line data checks are also made and fed back to the investigator.

As discussed earlier in Section 3, experience with earlier data collection programs indicates that the more steps involved (paper handling, etc.), the longer the delays. These delays have hampered the ability to perform follow-up investigations or to correct the data collection methodology. The data collection process has been completed (or nearly completed) before major areas of missing data or inconsistent interpretations have been discovered. Besides the need for timely analytic results, there is also the requirement for detecting potential problems before their solution becomes economically very costly. The central data collection program and the data system monitoring function are described in Sections 8 and 9, respectively.

## 6.2 Quick-Reaction System (QRS) Data System

The QRS element of NASS is intended to be addressed to a single specific highway safety problem topic at a time. Such topics may range from the evaluation of the performance of a new vehicle safety device being introduced into the population (for example a quick look at fuel systems in passenger cars built to the new FMVSS #301) to the determination of injury patterns associated with a defined (say, a side impact) type of crash. The data for a given QRS investigation might be a supplement for cases already collected in the CSS program, or might constitute a separate set of cases duplicating some of the data normally acquired in the CSS. In any case, the specific data form for collection of information under the QRS would be designed for that problem, distributed to the field agencies, and used for a prescribed period of time.

Timeliness is a critical aspect of the QRS, as the name implies. Consequently, the interactive data entry procedure has been designed as a generalized program that is independent of the specific set of questions asked. This permits the system manager to implement a new set of QRS data elements within days of their selection. The field investigator uses the same data entry protocol used in the CSS. The data entry program is self-documenting, so that the field investigator can interactively refresh his memory concerning any specific QRS data element he is unsure of.

Because the QRS operates in the environment of an ongoing CSS sample frame and set of PSUs, and because the data system is generalized for responsiveness, the Quick Response System is designed to be just what its name implies.

### 6.3 Sampling Frame Data System

The data entry process developed for the CSS and QRS in the previous sections is equally useful for entering the sample frame data. The specific data elements are those entered from the police report that are necessary for the selection of specific cases within the sample frame, i.e., accident identification numbers and sample selection data.

It is anticipated that each PSU will maintain manual logs of candidate accidents in the sample frame and of cases sampled for field investigation. Basic accident identification and sampling data elements will also be entered via the interactive computer terminals. This is done for two reasons: (1) the frame file is useful in itself for analysis and projection of the national accident population; and (2) by entering the frame data and matching its variables against the case selection criteria, the process of case selection from the frame can be automated and/or monitored.

The sampling data elements are of two types: sample class or strata, and random element. These will be used as the criteria for individual case selection within the sample frame. The sample class or strata data elements include such items as traffic unit type (car, truck, etc.), vehicle model year, hospitalization, and accident type.

Certain subclasses/strata will be sampled at a rate of less than 100%. For example, accidents in which no one was hospitalized might be sampled at the rate of 20%. The random data elements will be used to select cases within each subclass according to the prescribed sampling fraction for that subclass. These random data elements may be, for example, predetermined digits of certain police accident report numbers which could be expected to vary from form to form in a random fashion. Examples include portions of the police accident identification number, vehicle license plate number, or VIN.

The selection rules are based solely upon data elements entered into the sample frame file. The data entry program, in turn, uses these data elements to provide or confirm the selection of each case. A selection check routine has been written as a separate program module that is easily modified in response to changes in sample selection criteria.

## 6.4 Multidisciplinary Accident Investigation Data System (MDAIS)

The data system for the existing MDAI case studies has been using an annotated and augmented version of the Revision-3 Collision Performance and Injury Report (CPIR) form. The original Revision-3 CPIR was implemented in 1969, and has been continuously "patched" since that time. Consequently, many inefficiencies and redundancies have crept in.

The MDAI case studies should continue to be automated under the NASS program, either in the existing format or in a newer and more functional format. Since the MDAI cases are not part of the proper statistical sampling protocol used in NASS, the MDAI cases can be automated and analyzed as a completely separate function. The current MDAI case automation program is described elsewhere.*

*Multidisciplinary Accident Investigation Data File, 1974 Final Report, J. Marsh, Highway Safety Research Institute, Contract No. DOT-HS-4-0089. March, 1975, 133 pages.

# 7. FIELD DATA COLLECTION

## 7.1 Field Data Sources

The set of CSS data elements to be collected during the field accident investigation process was developed according to the approach described in the previous section (Section 6 - Data System Design). One of the criteria for selection of a field data element was that it indeed was collectible. It is most important that, once defined, a concerted effort be made to acquire that data element.

The CSS data elements discussed here were selected in part because they were generally obtainable from one or more of five sources: (1) police reports, (2) vehicle inspection, (3) driver or occupant interviews, (4) hospital records, or (5) site visits. There are, however, many other sources of information which may substitute or help to confirm information normally acquired from one of these. The investigator should not have the attitude that he should fill in the data element only if it is readily available, but rather that he should achieve as little missing data as possible. Table 7.2 indicates possible sources of information, all of which should be considered available to the investigator. A large percentage of the accidents (say 75%) may never require anything other than the police report, driver, vehicle inspection, and site visit--but 25% missing data in the system cannot be tolerated, and these other sources should be continually developed. To paraphrase Mr. Kennedy, "Ask not how much information you can get from the five principal sources, but where you have to go to get the information necessary to complete the forms."

## TABLE 7.1. CSS FIELD DATA SOURCES

(1) Police Report

- Standard Police Report
- Supplemental Investigative Reports
- Police Photographs

(2) Vehicle Inspection

(3) Interviews

- Driver
- Passengers
- Family/Friends
- Witness(es)
- Police, EMS, Fire, Towing Personnel

(4) Medical Reports

- Hospital Records
- Physicians Data
- Medical Examiner/Morgue/Autopsy Reports
- Death Certificates
- EMS Records

(5) Site Examination

(6) Administrative Records

- Vehicle Registration Records
- State Driver Records
- Court Records

(7) Local/Miscellaneous Records

- Weather Reports
- Traffic Records
- Road Comm. Reports
- Auto Repair/Service Records
- Fire Records

(8) Reference Information

- Motor Vehicle Identification Manual (VINs) National Auto Theft Bureau
- Vehicle Weight Data
- Highway Inventory Data

Data elements* planned for the CSS are listed below in general form associated with the principal source.

Data Obtained Primarily from Police Reports**

Accident Data
Time
Report Identification
Type of Accident (general type, i.e., collision with fixed object, rollover, sideswipe, etc.)
Location of Accident
Vehicles, Cyclists, Pedestrians Involved
- Vehicle Make
- Vehicle Year
- Vehicle Type

Vehicle Disposition (towed, where)
Driver Identification
- Age
- Sex
- Injury (police code)

Injured Occupant Identification
- Age
- Sex
- Injury (police code)

Hazardous Actions/Citations
Weather/Surface Conditions (dry, wet, snow, ice)
Lighting (daylight, darkness, dusk/dawn)

Data Obtained Primarily from Inspection of the Vehicle

Vehicle Identification/Description
- Model
- Style
- Function
- Cargo
- Weight/Loading

Damage Severity
- CDC
- Crush (inches)
- Damage Areas

*Data element is used here to describe a particular category of data, such as damage severity. The assessment of damage severity may require many computer variables to contain the data, which is the case with the present use of the Collision Damage Classification (CDC).

**These do vary from state-to-state. Those included here we felt to be most representative of what is generally found in most police accident reporting formats.

Damage to Components

- Door (jammed, latches, side door integrity)
- Hood (separation, release, penetration)
- Windshield (cracked, broken, separated)

Internal Damage/Contact Points
Restraint Equipment Available/Condition
Fuel Leakage (origin)
Fire (extent, origin)
Vehicle Speed Estimates

Data Obtained Primarily from the Driver/Operator/ Pedestrian/Occupant or Hospital Report

Seated Location
Weight
Height
Sex
Age
Posture
Injuries (AIS, OIC)
Injury Source/Contacts
Treatment
Ejection
Entrapment
Trip Origin/Destination/Purpose
Route/Area Familiarity
Vehicle Familiarity
Driver Education
Driving Experience (years, miles/year)
Impairment (physiological/pharamological/ psychological)
Violations/Citations

Roadway/Site Visit

Roadway Type/Traffic Lanes
Roadway Design Involvement (grade, superelevation, crown)
Roadside Involvement (median, roadside structures and hazards, maintenance)
Fixed-Obstacle Involvement (distance from roadway, type size, vehicle interaction)

## 7.2 Field Data Forms

As discussed earlier, to achieve periodic and timely summary data from NASS, the data handling system has been designed to provide for quick and interactive data entry from each PSU into the Data Center. While this has been established as the primary method of data flow, the investigator's field accident report form will serve as a back-up to the automatic remote entry system. A field form has been designed for use in the CSS of NASS and is included in this report as Appendix F.

The CSS field form is modular and is made up of five basic modules. These were designed to serve as a hard-copy back-up report, a convenient form for field use with "scratch pad" features, and for use as a ready reference to facilitate entry of accident data into the automatic entry system. The five modules of the field form are:

"A" Accident Module, a single-page form for accident descriptive and administrative data, accident sample selection criteria, and assessed causal factors.

"V" Vehicle Module, a two-sheet form for recording data regarding vehicle identification and description, accident damage, tire condition, specific vehicle component damage, internal damage, restraint features, and vehicle speed estimates, for use primarily during a case vehicle inspection.

"O" Operator Module, a one-page form for recording occupant seating, driver (or cyclist, pedestrian) physical description, injuries, restraint usage, impairment, citiations, experience and trip details; for use primarily during an interview.

"P" Passenger Module, a one-page form to record data on passengers in terms of physical characteristics, injuries, restraint usage, treatment, ejection or entrapment, also used during an interview.

"E" Environment Module, a one-page form for recording data regarding roadway and roadside features during a site visit.

Each module will be used as many times as needed. Only one Accident module would be used. Both Vehicle and Operator modules would be used for each traffic unit involved in the accident. The Passenger module would be used only when occupants other than the driver are involved. Up to two passengers can be recorded on each module, with more copies used, as required, to cover all the passengers.

Each module is tagged in the upper left corner with an identifying mnemonic letter. Space is also provided for encoding the accident case log number and vehicle number in the corner of each module. Thus the individual modules can be handled separately or readily combined. The organization or flow of questions within each module is structured to match the expected investigative protocol. For example, the Operator module follows the outline of a typical interview sequence. The specific questions are intended to be cues to the interviewer. They are not intended to be recited verbatim or in a strict sequence. The approach was to structure each module around the typical investigative protocol rather than in the format of the subsequent data analysis files. The organization of each module is therefore subject to alteration as a result of the pilot test.

Each of the data elements on the field forms follows one of two styles. If the responses are few and routine, (e.g., Y N ?) then all the responses are explicitly displayed, so that the investigator can simply circle or mark the appropriate response. Otherwise the responses are left as an open line on which the investigator may record his field notes (e.g., fire origin).

The completed forms serve as the investigator's field notes for inputting data to the computer system. Alternatively, the form can be forwarded to a data editor for encoding into a digital file. It is expected that the field investigator would retain a file copy of the field form, supplemented by a police report form, for future reference to that accident.

## 7.3 Field Data Coding Manual

One of the major problems that the CSS is expected to solve is the consistency of reporting in different regions of the U.S. The need for training and for monitoring for quality control has been discussed elsewhere in this report. The field investigators would also have a manual of instructions as a guide to their reporting efforts--basically a field investigators handbook which provides an in-depth explanation of the complete field investigation protocol including--case selection rules, investigation techniques, definition and interpretation of required data elements, instructions for data entry procedures and basic reference information. The handbook serves to complement the training and monitoring activities.

A prototype of a field data coding manual has been developed as one component of the investigators handbook (Appendix G ). The coding manual documents the interface between the investigator and computer file on a detailed data element level. It contains all of the valid responses for each data element. These responses are the same ones incorporated in the interactive data collection program described in the next section. This program, besides prompting for valid responses, also permits the investigator to display all the appropriate responses. In essence the printed coding manual is also stored on-line for investigators purusal.

The CSS field data coding manual is organized around the five modular forms. The upper corner of each page contains the module's identifying letter. The data elements follow the same sequence as the field forms and are numbered within each module, for easy reference. Thus A-17 refers to "First Harmful Event", the seventeenth question in the Accident module.

An effort was made to minimize the amount of code memorization or lookup required to enter a case and to make the transcription in the computer as natural as practical. For most of the data elements the investigator can enter an English "word" (e.g., "Bus") or a mnemonic truncation (e.g., "B" for "Bus"). Numeric responses are also entered in their natural form (e.g., "8/21/75" or 5'7").

Three standard abbreviated responses have been programmed as universal for all data elements:

| | |
|---|---|
| UNK or ? | for "Unknown stated by investigator" |
| N/A or / | for "Not/Applicable" |
| * | for Default value |

These responses are equally acceptable for both "word" and numeric responses. Thus "?" is a valid response for both First Harmful Event and Accident Time. These standard abbreviated responses are not permitted where such a response is invalid. For example, "?" is not permitted for Accident case ID Number and "N/A" is not permitted for Accident Date. The default response ("*" or a carriage return) can be used to answer one question or an entire set of questions with a predetermined response. Thus the investigator can readily default questions concerning Axles, Tread and Door Latches Separated to "Not Applicable" when the traffic unit being reported is a pedestrian.

For data elements with a large number of valid responses (say over 25) the mnemonic or natural responses become so long that they are hard to remember and are difficult to enter correctly. For example the data element "First Harmful Event" required a minimum of a four letter mnemonic in order to provide unique responses, so both a numeric code or a "word" depending on the investigator's preference (e.g., either "12" or "Fire" are valid). For longer lists (e.g., Causation Factors) only numeric codes are permitted.

# 8. CENTRAL DATA COLLECTION

## 8.1 CSS Data

The CSS field data forms described in the preceding section are used as a convenience by the investigation teams to ensure that all the data elements are available at the time of input to the system. The input is done directly from the field team location to the central computer. This recommendation will be tested during the pilot phase of NASS opearation by comparing the direct digital input against a "paper" system, but at present we think that digital input will be not only more efficient in terms of speed and accuracy, but, in the long run, less expensive to operate.

Using the completed field data form for reference, the investigator communicates with the system via a data entry terminal over ordinary telephone lines. The ENTRY program as shown in Figure 8.1 performs these entry functions, and examples of the program operation appear in Appendix B. The program itself is reproduced in Appendix B. It is written in Fortran IV with assembly language input/output routines, operating within the MTS (Michigan Terminal System) Operating System on an IBM 370/168. The ENTRY program functions are detailed in Figure 8.2.

The entering of data is made in an interactive mode, with the operator being queued by the program for each of the data elements to ensure, as nearly as possible, complete and accurate entry of a case with a single pass. Upon signon, the operator may set a series of switches, determined largely by his own

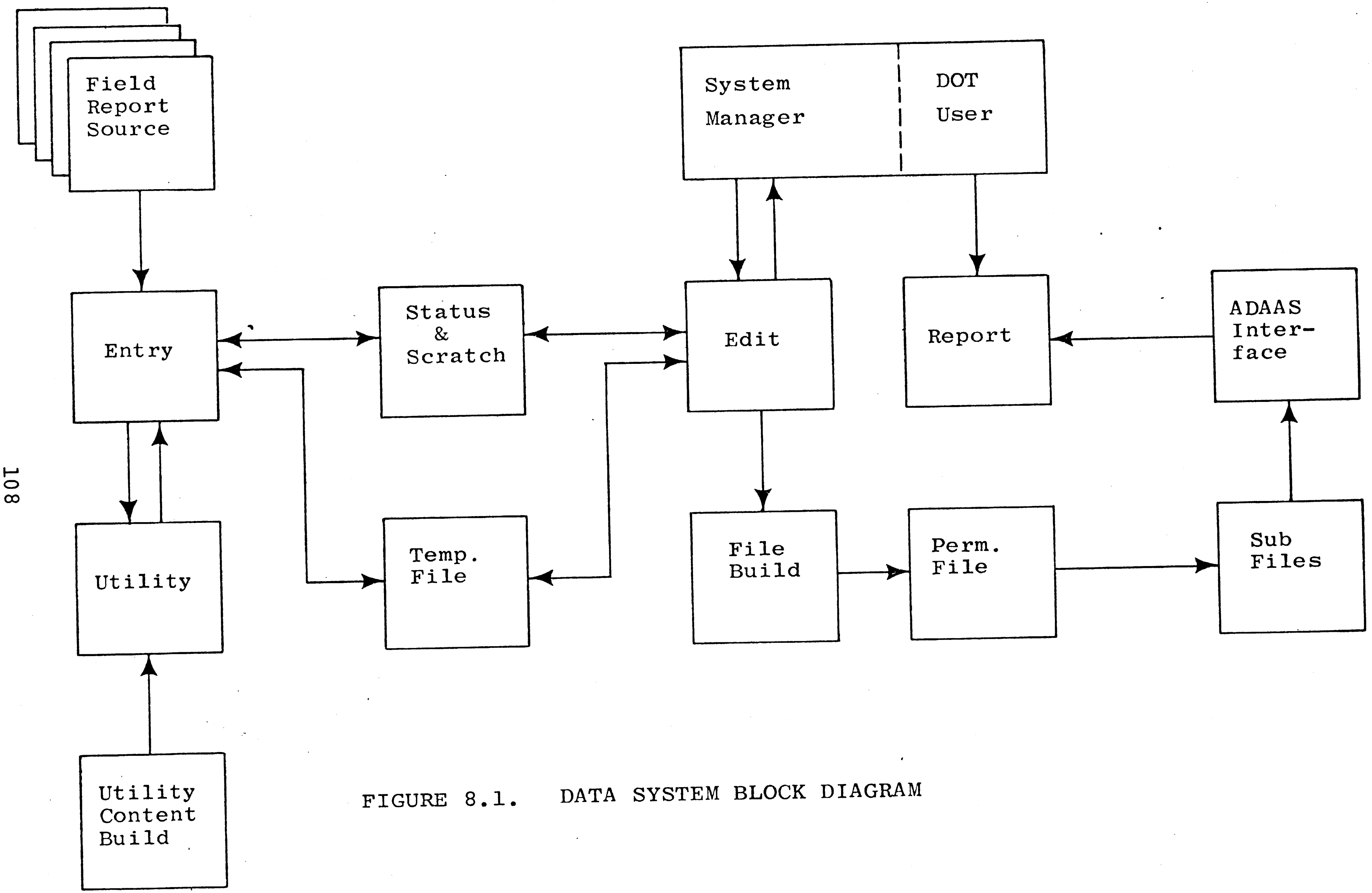


FIGURE 8.1. DATA SYSTEM BLOCK DIAGRAM

Switches

- Long/Short Form
- Echo On/Off
- Group Yes/No

Content

- Variables
- Dictionary
- Permissible Values
- Error Messages

Process

- Ordering
- Describe Command
- Grouping
- Scratch Pad Interaction
- Filing
- Feedback & Revision

FIGURE 8.2. ENTRY PROGRAM FUNCTIONS

preference and experience with the system. The long form spells out the data elements completely, while the short form presents only variable numbers and abbreviated forms of the data element labels. The Echo or repeating of the input by the computer may be requested or supressed, and the data elements may be entered singly or in groups.

The ENTRY program accesses a list of variables containing the data elements required as input. This dictionary is used by the system manager in defining the entries which can be made and for use in cross-checking on the compatibility of entries with one another. Permissible values are stored for each required entry in the form of either numerical limits for entries or a list of descriptive terms allowed. The program has been designed so that the list of permissible values may be changed during the course of the NASS program. Any English or numerical phase may be entered and will be accepted in its literal form. It will be tagged as a "new" value for consideration by the system manager at the data collection center and can then either be added to the list of permissible values or the entering user will be requested to use some other term instead. Error messages form a part of the program associated with each variable, and are used to cue the investigator in an attempt to get clean case input on the first pass.

The process functions of the ENTRY programs do the bookkeeping necessary for system operation. The ordering portion of the program keeps track of the entries, and the describe command is available for a user who is unfamiliar with or may have forgotten what is required in connection with a particular variable. The grouping sub-routine allows the entry of clusters of data elements. This permits faster case preparation after users become familiar with the system. The scratch pad part of the program is designed to provide two-way communication between the investigator and the system manager for such information as entry and case status, and, as mentioned above, the addition of new values to data element variables. The filing portion of the ENTRY program prepares the data elements of a case for storage in the case files for later use by the system. This file is a temporary one, and is used to display the information in condensed form to check for errors and to revise when necessary.

Returning to Figure 8.1, we contrinue with the description of the entire program. Note that the box labeled STATUS & SCRATCH forms the only line of communication between the field data sources and the system managers. They may, and undoubtedly will, speak to each other on the telephone, but the intention of this design is to have the major portion of the administrative functions carried out internal to the computer.

The EDIT routine is the final check on data quality by the system manager before its entry into permanent storage in the data bank. Since the major portion of editing for inaccuracy in the data has already been performed at the initial entry, the primary function of this level of activity will be to reduce to as close to zero as possible the amount of missing data. No case will be entered into permanent storage until the manager is assured that no variable for a particular case has been omitted unnecessarily. The manager makes use of the scratch pad to communicate with the field data sources and to monitor the status of data collection throughout the system.

A case, having passed through the edit process and having been certified as complete, is entered into permanent storage. At full-system operation, involving 35 field data sources, each with a responsibility for 500 cases per year, the load on the editing portion of the system will average about 60 cases per day. At this time during the data handling process, new analysis variables are created from the data elements that are necessary or desirable for the output of the system. This process is performed routinely by the FILE BUILD program. For example, while the age of the driver is recorded and entered as a data element, the permanent data file will store, along with this value, the age of the driver coded in the appropriate NSC and five-year groupings.

The permanent master file is maintained as an occupant file, with the traffic unit and accident information for each case duplicated in the occupant master record. A schematic of this structure is shown in Figure 8.3. For the generation of reports, however, it would be inefficient to operate on the entire master file, and it will be routine to build sub files from it for that purpose. At this time it is thought that three analysis files will be constructed, dealing with accidents, vehicles, and occupants. These analysis files will be updated with new cases entered into permanent storage on a periodic basis, weekly or even daily, as the system approaches full-scale operation.

### 8.2 Quick-Reaction System (QRS) Data

The key to the operation of the QRS portion of the system is the UTILITY CONTENT BUILD subroutine indicated in Figure 8.1. This feature of the data handling system allows changes in the data elements to be made

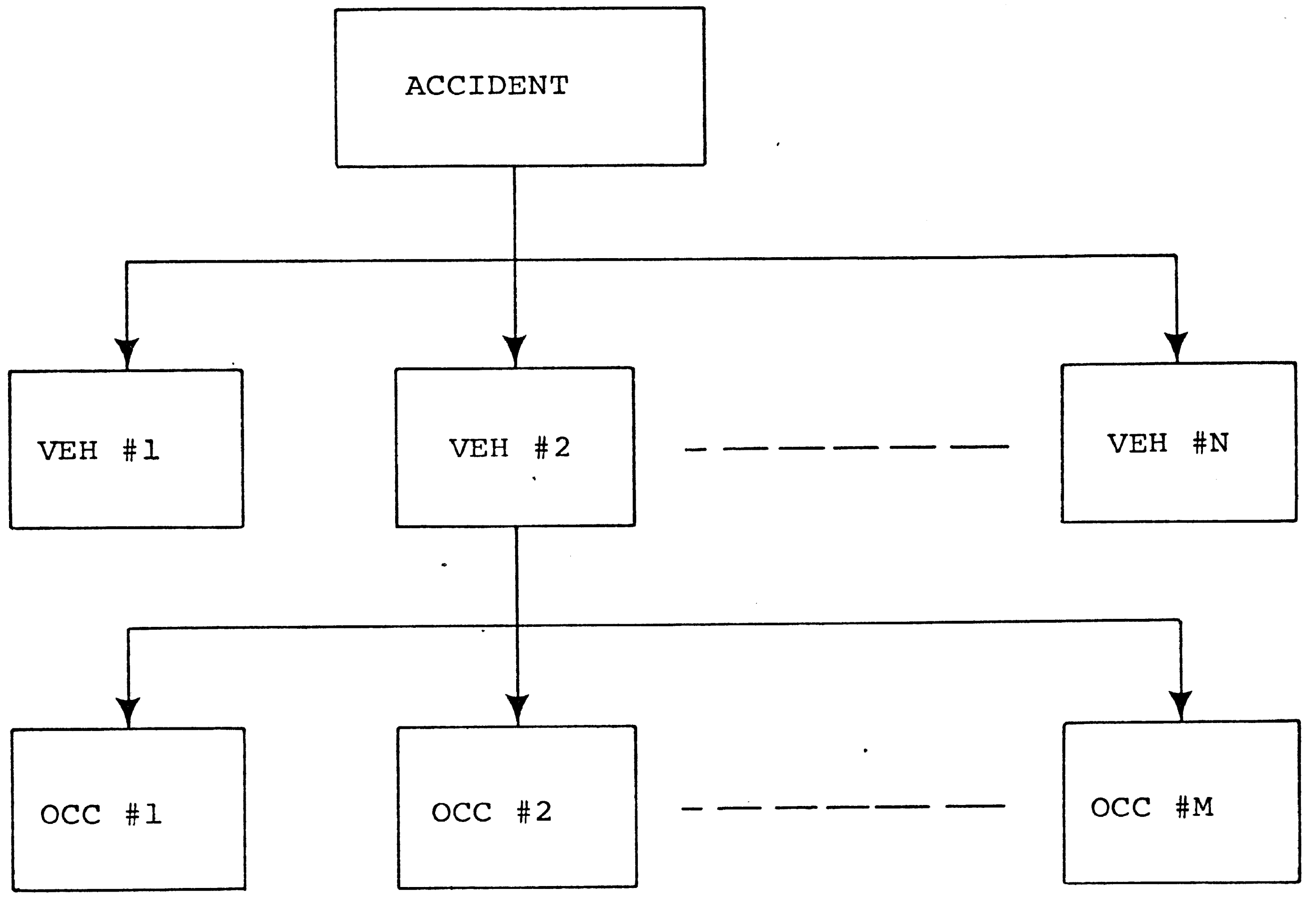


FIGURE 8.3. FILE STRUCTURE

quickly and easily. This program allows the construction of a UTILITY package to operate within the ENTRY program for any particular application. Once decided upon, the data elements, together with the limits on their values, can be prepared for system operation in a matter of hours. The interactive mode of the ENTRY program then operates on the new content; files are built surrounding the new data elements; and the output routines can be run against these files.

With minor modifications to the RULE subroutine, cases to be studied in a QRS or Special Study operation of the system may be identified from the sample frame file. The QRS study can then be included in the status reporting and monitoring functions of the system, described in Section 9 of this report.

## 8.3 Sample Frame Data

The preceding sections described the central data collection process for the CSS and QRS investigations. These cases are selected from a sample frame of specific police-reported accidents. The sample frame data is a list of accidents and is also maintained as a computer file.

As each sample frame entry is concluded, the RULE program is evoked and the field team is supplied with immediate feedback as to whether the accident is to be (or should have been) selected as a case for CSS or QRS study. Should the computer not be available, the field teams will be provided with a manual backup for case selection.

The entry of frame data is done in exactly the same way as that of field cases. The entry program, however, calls for far fewer variables, and the entry process requires a single line for the accident variables and one line for each of the traffic units involved in it. The investigator initially selects the frame-made rather than the case-mode of program operation, before sample frame data are entered.

# 9. DATA SYSTEM MONITORING

One of the most important functions of the NASS operations is managing the data collection efforts of the field teams (Section 7) and their entering of data into computer storage (Section 8). Monitoring is done at the central data point by means of the status-keeping routine incorporated into the EDIT program, and at the field level through the scratch pad status and communication device. At the intermediate management level, the zone centers described in the next section, system status information is provided to allow the day-to-day determination of adequate conformance to system requirements by the teams under zone center control.

At a field location, every signon at a terminal device automatically calls up the status and scratch pad information to the user. Table 9.1 shows a sample printout of what may appear to the user. The first three lines show his progress with respect to his case-entering activities. The next three entries give case numbers for reference to the paper forms maintained at the field location, and they describe the disposition of cases as seen by the editor at the central site. The last entries have been made by the system manager based on his review of the cases entered. In this way the attention of the field team can be called to matters requiring immediate attention, so that the possibility of missing data remaining in the file is minimized.

The zone manager has responsibility for ensuring the smooth operation of the field teams under his immediate control. He does this by providing

TABLE 9.1. FIELD TEAM STATUS PRINTOUT

```
$SIGNON XXX

ENTER USER PASWORD


                    USER XXX
                     STATUS

CASES IN PERM FILE TO DATE:  172
CASES IN TEMP FILE NOT COMPLETED:  12
CASES ASSIGNED, NOT YET ENTERED:  2

CASE NOS. COMPLETE, NOT YET EDITED:
     234, 237, 240, 241, 242, 244, 245, 247, 249
CASE NOS. WITH MISSING DATA:
     233, 235, 236
CASE NOS. NOT YET ENTERED:
     248, 250

CASE NO. 233!
     VARIABLE 27 MISSING, DRIVER AGE?
CASE NO. 235!
     VIN NUMBER AND BODY STYLE DO NOT AGREE
     PLEASE CONFIRM
CASE NO. 236!
     MISSING VARIABLES 36, 38, 39 and 41.
```

professional personnel assistance when needed, by interpreting the system requirements where they may become unclear, and by anticipating possible problems in the data collection and entry process. To help him in this last function he may call up a set of status displays that both describe the condition of an individual team's operation and compare the operations of the teams with each other.

Table 9.2 is an example of the status of a single user which may be called by the zone manager. The frame and case files are monitored to arrive at some insight as to whether accidents are being identified at the rate at which they were expected to occur, and whether the investigations are proceeding at an acceptable pace. In the first case, a divergence between the expected and actual number of cases might indicate a change in police reporting procedures requiring alternate methods of accident identification. The second display line gives an indication of possible staffing problems or delays in completing the data entry.

The individual cases can be reviewed, as in the next portion of the display, to see how far behind in time the input has fallen and to discover any data elements which may be causing particular trouble requiring an upgrading of the team's professional expertise. The last table displayed is to alert the zone manager of any unusual delays or high priority items which might require assistance from the manager's staff.

Table 9.3 presents an example of the summary information the zonal manager might call regarding all the field teams under his control. This display would be used for comparative purposes in determining where managerial attention could be most profitably directed.

TABLE 9.2. INDIVIDUAL TEAM STATUS

TEAM XXX STATUS

14 AUG 77

| FRAME: | EXPECTED | ACTUAL | A/E(%) |
|---|---|---|---|
| | 385 | 370 | 96 |

| CASES: | ASSIGNED | PERM | TEMP | NO ENTRY |
|---|---|---|---|---|
| | 125 | 109 | 10 | 4 |

MISSING DATA CASES:

| CASE NO. | DATE | MISSING VARIABLES |
|---|---|---|
| 324 | 9 Jul | 27, 29 |
| 327 | 12 Jul | 32 |
| 334 | 15 Jul | 27, 32 |
| 335 | 15 Jul | 44, 45, 46 |
| 340 | 19 Jul | 27 |
| 343 | 26 Jul | 4, 27, 32 |
| 345 | 30 Jul | 27, 32, 44 |
| 354 | 4 Aug | 4, 8, 27, 32 |
| 364 | 12 Aug | 4, 27, 29, 32, 44 |
| 365 | 12 Aug | 8, 27, 29, 32, 44 |

NO ENTRY CASES:

| CASE NO. | DATE | MOD. YR. | INJURY |
|---|---|---|---|
| 366 | 13 Aug | 75 | 0 |
| 367 | 13 Aug | 76 | 3 |
| 368 | 13 Aug | 74 | 0 |
| 369 | 14 Aug | 77 | 6 |

TABLE 9.3. ZONE STATUS
14 AUG 77

| | TEAM | | | |
|---|---|---|---|---|
| | XX1 | XX2 | XX3 | XX4 |
| FRAME: | | | | |
| Expected | 385 | 395 | 350 | 370 |
| Actual | 370 | 400 | 340 | 330 |
| A/E (%) | 96 | 103 | 97 | 89 |
| CASES: | | | | |
| Assigned | 125 | 129 | 116 | 120 |
| Perm | 109 | 120 | 83 | 105 |
| Temp | 10 | 9 | 17 | 13 |
| Miss. Data | 10 | 2 | 15 | 7 |
| No Entry | 4 | 0 | 16 | 2 |
| FATALS: | 6 | 8 | 3 | 4 |

At the highest level of system management, the data center, complete access to all information contained in the files is available. Individual style of management and information requirements will play an important role in determining how the monitoring function is performed. Furthermore, the computer programming capability at the data center will be such that the managerial needs for monitoring displays can be easily met as they develop. The following tables should, therefore, be considered only suggestive of what may be produced at this level.

Table 9.4 indicates the overall status of the NASS with respect to zonal activities. Similar data could be requested for each zone broken out by team to aid the zone managers in their control over the system operation. The manager at the data central might, on the basis of this display, check further into the details of zone 4, to determine which of the teams is mainly responsible for the inordinately high missing data rate in the CSS portion of the system.

A more sophisticated level of monitoring is provided via the analysis capabilities described in Section 10. Analytical runs can be routinely performed to test for differences or changes in the means or distributions of individual variables. These differences should be tested across time and on an inter-team basis. Significant differences between teams may be indicative of inconsistent investigation or coding practices, and a need for further training and guidance. For example, the finding of a more than 2-to-1 differences in the frequency of AIS-2 lacerations (as in the 73/74 restraint study) may dictate a thorough review of AIS coding practices by all teams.

TABLE 9.4. DATA CENTER STATUS DISPLAY
SYSTEM STATUS
14 AUG 77

| | ZONE 1 | ZONE 2 | ZONE 3 | ZONE 4 | ZONE 5 |
|---|---|---|---|---|---|
| TEAMS | | | | | |
| ACTIVE | 3 | 3 | 5 | 4 | 4 |
| AUTHORIZED | 6 | 6 | 7 | 7 | 6 |
| PERSONNEL | | | | | |
| AUTHORIZED | 14 | 15 | 22 | 19 | 21 |
| ACTIVE | 14 | 14 | 20 | 18 | 21 |
| MDAI CASES | | | | | |
| EXPECTED | 20 | 22 | 36 | 24 | 22 |
| COMPLETE | 18 | 20 | 35 | 26 | 18 |
| CSS CASES | | | | | |
| ASSIGNED | 730 | 720 | 806 | 696 | 704 |
| FILED | 705 | 715 | 792 | 685 | 688 |
| QRS STUDY #5 | | | | | |
| EXPECTED | 88 | 92 | 116 | 98 | 112 |
| FILED | 76 | 104 | 92 | 92 | 116 |
| MISSING DATA RATE | | | | | |
| MDAI (%) | 0.8 | 1.1 | 0.9 | 0.6 | 0.9 |
| CSS (%) | 1.0 | 0.8 | 3.4 | 6.1 | 1.1 |
| QRS (%) | 2.0 | 0.8 | 1.2 | 1.4 | 2.4 |

# 10. DATA ANALYSIS AND REPORTING

As stated in Section 3 of this report, the purposes of NASS are to provide estimates of national rates and trends with regard to the accident population. Each case reported into the permanent file of the system contributes toward the national estimate in an amount corresponding to its weight derived from the sampling fractions and probabilities as described in Section 4.4. The interrogation of the data file will yield at any time a description of accident rates, and, across time, will provide estimates of trends. This section provides a description of how data anlysis and reporting fits into the overall system design.

## 10.1 Analysis File Structure

The cases reported to the central computing facility by the field data collection teams are verified and entered into permanent storage in. a master file. This file is an "occupant" file, in that a complete case record is constructed for each occupant (or pedestrian, bicyclist, motorcyclist) in every accident. The case records for occupants in the same vehicle are redundant with respect to the vehicle information, and for vehicles in the same accident are redundant in their accident information. This master file will rarely, if ever, be used directly in the performance of analytical tasks. From the master file, analysis files will be created in order to facilitate use of analysis tools. Based on the kinds of questions that are expected to be asked of the data system, three analysis files have been detailed: (1) Accident, (2) Traffic Unit, and (3) Occupant.

### 10.1.1 Accident File

The accident file records variables specifically concerned with the nature and consequences of the traffic accident. These data are oriented toward the events surrounding the accident, the environmental conditions at the time of occurrence, and the physical situation prevailing at the time. Also detailed are variables concerned with the accident configuration and any inferences which have been made in the investigation with regard to the causal factors involved. Incorporated into this file are summaries of the vehicle and occupant injury information from the master file insofar as they relate to the accident as a whole. Such data elements as the total number of vehicles or traffic units involved, the most severely damaged vehicle, the worst injury to any occupant, and the disposition of vehicles and victims are generated from the master file for use in subsequent accident analyses.

### 10.1.2 Traffic Unit File

This file is constructed for the specific purpose of determining what happens to specific vehicles in an accident. Vehicle parameters such as weight, size, and type are present in the file along with the accident specifics of speed, direction of motion, and the object struck in the collision. These variables can be compared to the amount of damage done both to the vehicles and their occupants, and can be related to defects predating the crash, the type and use of safety equipment, and driver-related factors contributing to the crash and the resulting damage.

### 10.1.3 Occupant File

The variables maintained in this file are primarily related to the occupant and his/her condition prior to, during, and subsequent to the crash. The type and amount of injury are capable of relation to the accident and vehicle variables relevant to the circumstances as well as to other parameters of the occupant such as age, sex, and physical characteristics. Other important data elements in this file have to do with the seated position of the occupant, the use of restraining devices, and the interior portions of the vehicle which were involved in causing injury.

## 10.2 Analysis

The cumulative data in the analysis files represent a continual updating of estimates (together with the appropriate sampling errors) of statistics regarding national accident phenomena. This is because, as described in Section 4.4 of this report, each case comes into the data file with its own weight for the reconstitution of the national estimate. It remains, therefore, only to extract these data in a useful form for purposes of analysis.

The validity of this process has been tested at HSRI during the past year in connection with the Restraint System Study supported jointly by NHTSA and the Motor Vehicle Manufacturers Association. This preliminary study has shown that the concept works well, not only in the analysis area, but also in the matter of report generation, discussed in the next section.

The following tables were generated as examples of NASS analysis statistics. The assumption underlying these is that 35 teams would be operating in 35

Primary Sampling Units (PSUs) throughout the nation. This assumption is arbitrary but is based on a data collection and subsequent analysis cost of $6 million per year. The PSUs will be selected so that the smallest of them will yield 3,000 accidents annually, and secondary sampling will take place in those where a larger number is expected. This will generate in our sample 90K accidents per year, comprising the accident records which will be maintained and from which the CSS cases will be selected.

## 10.3 Report Generation

The user will have access to the NASS data files through existing data analysis program packages such as SPSS, OSIRIS, and ADAAS.* These systems have been in use a number of years and provide the analyst a straightforward method for recovering data from a large file. In the HSRI Automated Data Access and Analysis System (ADAAS), the analyst can readily access a variety of analysis programs and data files on a keyword basis. The output is organized, formatted, and labelled automatically.

As easy as ADAAS is to run, the process was further simplified for use in the Restraint System Study. The commands used to initiate the ADAAS run are normally

*"SPSS - Statistical Package for the Social Sciences," N. Nie, D. Bent, C. Hull. McGraw-Hill Book Co., 1970, 343 pages.
"OSIRIS III" - Volumes 1 to 6, Center for Political Studies, Institute for Social Research, Ann Arbor, Michigan, 1973.
"A Manual for ADAAS", Preliminary Version, Highway Safety Research Institute, Ann Arbor, Michigan, February, 1974, 111 pages.

entered by the analyst from a terminal, on-line with the computer. In this case, however, these commands were stored in a computer file, and the entire series of reports could be produced by issuing a single instruction which accessed that file. A sample of the output produced by this means is shown in Table 10.1. The updating of the data file would, of course, yield a table in the same format but containing different values. This procedure will be used in the generation of routine reports for NASS. The desired reports will be designed using ADAAS commands, and then stored in a computer file along with an identification number. The analyst must simply indicate which specific report he desires and it will be automatically produced and will be based on the data currently contained in the NASS data file.

An important output of the NASS will be the supplying of analyst with trend information. Computer techniques are available to plot curves from data in digital form, and these will be interfaced with the report generation portion of the system. Figure 10.1 represents an example of such a plot, showing the mean driver age (separated into male and female) over time. Reports of this type can be called by number or, if the analyst wishes less than the full file content, he may specify, in an interactive mode, the limits he would like displayed. The selection of cases for detailed analysis under NASS will consist of 500 vehicles within each PSU. The resulting NASS files will contain, for each year of data collection at full-scale operation, 15,000 vehicles and approximately 25,000 occupants. The sample N's and percentages were estimated using the data from the 1973 Texas 5% sample file.

A. N and % in national sample of 15,000 vehicles.

I. Specific Vehicle Type

| | N | % |
|---|---|---|
| Passenger Car | 12,060 | 80.4 |
| Truck | 2,313 | 15.4 |
| Truck + | 404 | 2.7 |
| School Bus | 20 | .1 |
| Bus | 33 | .2 |
| Motorcycle | 195 | 1.3 |
| Pedestrian | 110 | .7 |

II. Most Serious Injury in Vehicle

| | N | % |
|---|---|---|
| No Injuries | 19,521 | 86.8 |
| K | 93 | .4 |
| A | 569 | 2.5 |
| B | 1,329 | 5.9 |
| C | 988 | 4.4 |

III. Day of Week

| | N | % |
|---|---|---|
| Sunday | 1,662 | 11.1 |
| Monday | 2,069 | 13.8 |
| Tuesday | 2,011 | 13.4 |
| Wednesday | 2,033 | 13.6 |
| Thursday | 2,061 | 13.7 |
| Friday | 2,712 | 18.1 |
| Saturday | 2,441 | 16.3 |

## IV. Driver Age NSC Groups

| | N | % |
|---|---|---|
| 0-14 | 146 | .97 |
| 15 | 316 | 2.11 |
| 16 | 614 | 4.09 |
| 17 | 718 | 4.79 |
| 18-19 | 1,411 | 9.41 |
| 20-24 | 2,768 | 18.45 |
| 25-34 | 3,366 | 22.44 |
| 35-44 | 2,090 | 13.93 |
| 45-54 | 1,673 | 11.15 |
| 55-64 | 1,070 | 7.13 |
| 65-74 | 626 | 4.17 |
| 75+ | 203 | 1.35 |

## V. Driver Sex

| | N | % |
|---|---|---|
| Male | 10,297 | 68.6 |
| Female | 4,703 | 31.4 |

## VI. Damage Scale (TAD)

| | N | % |
|---|---|---|
| 0 | 429 | 2.9 |
| 1 | 5,912 | 39.4 |
| 2 | 4,188 | 27.9 |
| 3 | 2,805 | 18.7 |
| 4 | 905 | 6.0 |
| 5 | 344 | 2.3 |
| 6 | 239 | 1.6 |
| 7 | 176 | 1.2 |

VII. Accident Type

| | N | % |
|---|---|---|
| Pedestrian | 170 | 1.1 |
| Other Vehicle | 10,719 | 71.5 |
| Railroad Train | 45 | .3 |
| Parked Car | 1,366 | 9.1 |
| Bicycle | 93 | .6 |
| Animal | 205 | 1.4 |
| Fixed Object | 1,283 | 8.6 |
| Other Object | 63 | .4 |
| Overturn in Road | 163 | 1.1 |
| Ran Off Road | 810 | 5.4 |
| Other Non-Collision | 84 | .6 |

VIII. Road Classification

| | N | % |
|---|---|---|
| Interstate | 1,520 | 10.1 |
| US & State Trunk | 4,409 | 29.4 |
| State II° Road | 863 | 5.8 |
| County Road | 419 | 2.8 |
| City Street | 7,764 | 51.8 |
| Turnpike | 25 | .2 |

IX. Age of Vehicle

| | N | % |
|---|---|---|
| 1 or less | 1,769 | 11.8 |
| 2 | 2,084 | 13.9 |
| 3 | 1,580 | 10.5 |
| 4 | 1,495 | 10.0 |
| 5 | 1,519 | 10.1 |
| 6-7 | 2,461 | 16.4 |
| 8-9 | 1,954 | 13.0 |
| 10-11 | 1,122 | 7.5 |
| 12 + | 1,016 | 6.8 |

X. Vehicle Damage Area

| | N | % |
|---|---|---|
| Front | 5,225 | 34.8 |
| Rear | 2,851 | 19.0 |
| Left | 3,286 | 21.9 |
| Right | 3,382 | 22.5 |
| Side & Top | 257 | 1.7 |

XI. Road Surface

| | N | % |
|---|---|---|
| Dry | 11,835 | 78.9 |
| Wet | 2,823 | 18.8 |
| Muddy | 10 | .07 |
| Snowy | 68 | .45 |
| Icy | 264 | 1.76 |

XII. Light Conditions

| | N | % |
|---|---|---|
| Daylight | 10,616 | 70.77 |
| Dawn | 69 | .46 |
| Dark - Not Street Lights | 3,141 | 20.94 |
| Dark - Street Lights | 884 | 5.89 |
| Dusk | 290 | 1.93 |

XIII. Vehicle Defects

| | N | % |
|---|---|---|
| None | 14,735 | 98.23 |
| Brakes | 150 | .99 |
| Steering | 15 | .10 |
| Lights | 7 | .04 |
| Windshield Wiper | --- | .00 |
| Tires | 52 | .34 |
| Trailer Equipment | 18 | .12 |
| Stop/Turn Signal | 9 | .06 |
| Wheel Came Off | 15 | .10 |

The following tables have been produced to show how this method of analysis can yield answers to a selected list of questions of concern to those who interrogate the NASS data files.

(1) What is % of accidents by injury level?

| Injury | None | K | A | B | C |
|---|---|---|---|---|---|
| % | 86.8 | .4 | 2.5 | 5.9 | 4.4 |
| N | 13,020 | 60 | 375 | 885 | 660 |

(2) Involvements by impact direction?

| Damage Area | Front | Rear | Left | Right | Side & Top |
|---|---|---|---|---|---|
| % | 34.8 | 19.0 | 21.9 | 22.5 | 1.7 |
| N | 5,220 | 2,850 | 3,285 | 3,375 | 225 |

(3) Percent of Accidents by severity and configuration that involve alcohol and drugs?

| | Injury | | | | |
|---|---|---|---|---|---|
| | None | K | A | B | C |
| Front | 125 | .6 | 4 | 9 | 6 |
| Rear | 69 | .3 | 2 | 5 | 3 |
| Left | 79 | .4 | 2 | 5 | 4 |
| Right | 81 | .4 | 2 | 6 | 4 |

Number of Alcohol-Related Involvements by Injury and Configuration for Citation.

|  | Injury |  |  |  |  |
|---|---|---|---|---|---|
|  | None | K | A | B | C |
| Front | 498 | 2 | 14 | 34 | 25 |
| Rear | 95 | 1 | 8 | 19 | 14 |
| Left | 109 | 1.5 | 9 | 21 | 16 |
| Right | 112 | 1.5 | 9 | 22 | 16 |

Same assuming 11% involvement.

(4) Percentage of accidents involving factor X.

| Factor | % | N Sample |
|---|---|---|
| Pick-up Trucks | 12.30 | 1,845 |
| Pedestrians | .77 | 116 |
| Motor Home/Camper | .066 | 10 |
| Motorcycle | 1.37 | 206 |

| | Injury | | | | |
|---|---|---|---|---|---|
| | None | K | A | B | C |
| Front | 498 | 2 | 14 | 34 | 25 |
| Rear | 95 | 1 | 8 | 19 | 14 |
| Left | 109 | 1.5 | 9 | 21 | 16 |
| Right | 112 | 1.5 | 9 | 22 | 16 |

Same assuming 11% involvement.

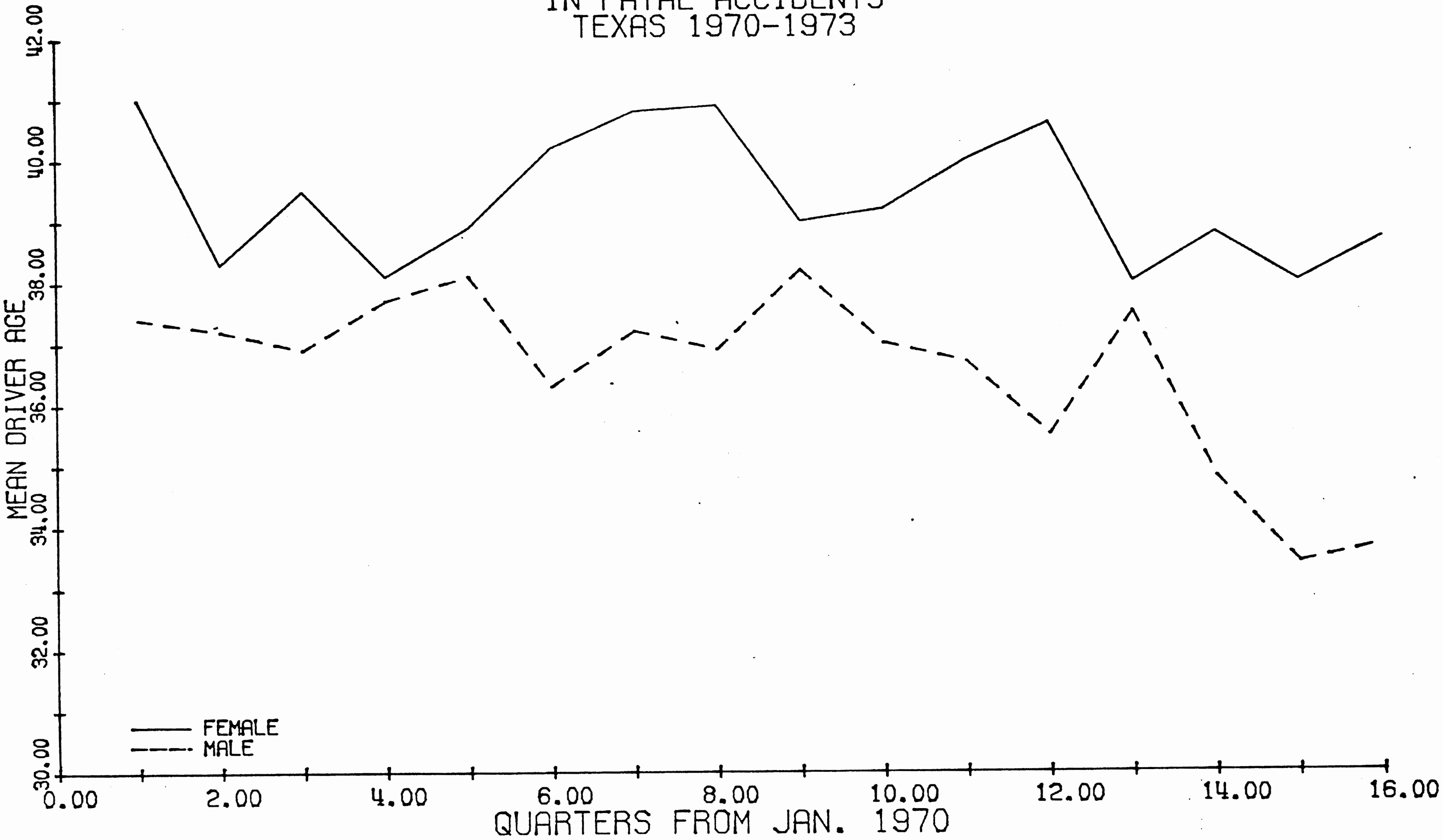
(4) Percentage of accidents involving factor X.

| Factor | % | N Sample |
|---|---|---|
| Pick-up Trucks | 12.30 | 1,845 |
| Pedestrians | .77 | 116 |
| Motor Home/Camper | .066 | 10 |
| Motorcycle | 1.37 | 206 |

TABLE 10.1. INJURY SEVERITY BY MODEL YEAR (Data weighted on inverse of sample fraction; occupants with known injury severity are omitted from Table.)

|  |  | AIS |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model Year |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | N |
| CALSPAN |  |  |  |  |  |  |  |  |  |  |  |  |
| 1973 | % | 47.5 | 39.2 | 9.3 | 2.8 | 0.4 | 0.2 | 0.2 | 0.2 | 0.2 | 0 | 495 |
| 1974 | % | 49.6 | 37.2 | 10.7 | 1.7 | 0.2 | 0 | 0.2 | 0.2 | 0 | 0 | 401 |
| HSRI |  |  |  |  |  |  |  |  |  |  |  |  |
| 1973 | % | 43.7 | 50.4 | 4.0 | 0.9 | 0.4 | 0 | 0.2 | 0.2 | 0.1 | 0.1 | 822 |
| 1974 | % | 44.5 | 48.8 | 4.5 | 1.6 | 0.1 | 0 | 0.3 | 0 | 0.1 | 0 | 692 |
| SwRI |  |  |  |  |  |  |  |  |  |  |  |  |
| 1973 | % | 37.3 | 47.6 | 8.3 | 1.1 | 0.3 | 0.2 | 0.1 | 0.2 | 0.3 | 0 | 1148 |
| 1974 | % | 49.8 | 42.7 | 5.9 | 1.0 | 0.3 | 0.1 | 0 | 0 | 0.1 | 0 | 864 |

FIGURE 10.1
TREND IN MEAN DRIVER AGE
IN FATAL ACCIDENTS
TEXAS 1970-1973



# 11. ORGANIZATIONAL COMPONENTS AND THEIR FUNCTIONS

The major functions of the National Accident Sampling System have been addressed in Sections 6 through 10. They include data collection (for the CSS, QRS, and MDAI activities); data handling, analysis, and reporting; command and control of the system; and monitoring of operations. In this section we consider alternatives for actual operation of the system, given that the above functions are to be performed.

Figure 11.1 is a functional block diagram of the system. Solid lines indicate the flow of data in raw or processed form, proceeding from the collection to processing and storage to the analysis and routine statistics production functions. Output of the system is directed to users at NHTSA and elsewhere through a management function. Command and control activities are shown by the dotted lines. A training function assures that personnel in the system are operating consistently on a long-term basis, and the quality control function serves the same purpose on a shorter-term basis. Direction of those activities in the system which change with time (e.g., new MDAI assignments, or new QRS assignments) is accounted for by a planning function as shown.

There are several ways of assigning the various functions to a working organization. Three arrangements will be discussed here, identified as a centralized, a decentralized, and a hybrid system arrangement. Although there are examples of each which are working systems, the hybrid system will be recommended for the NASS.

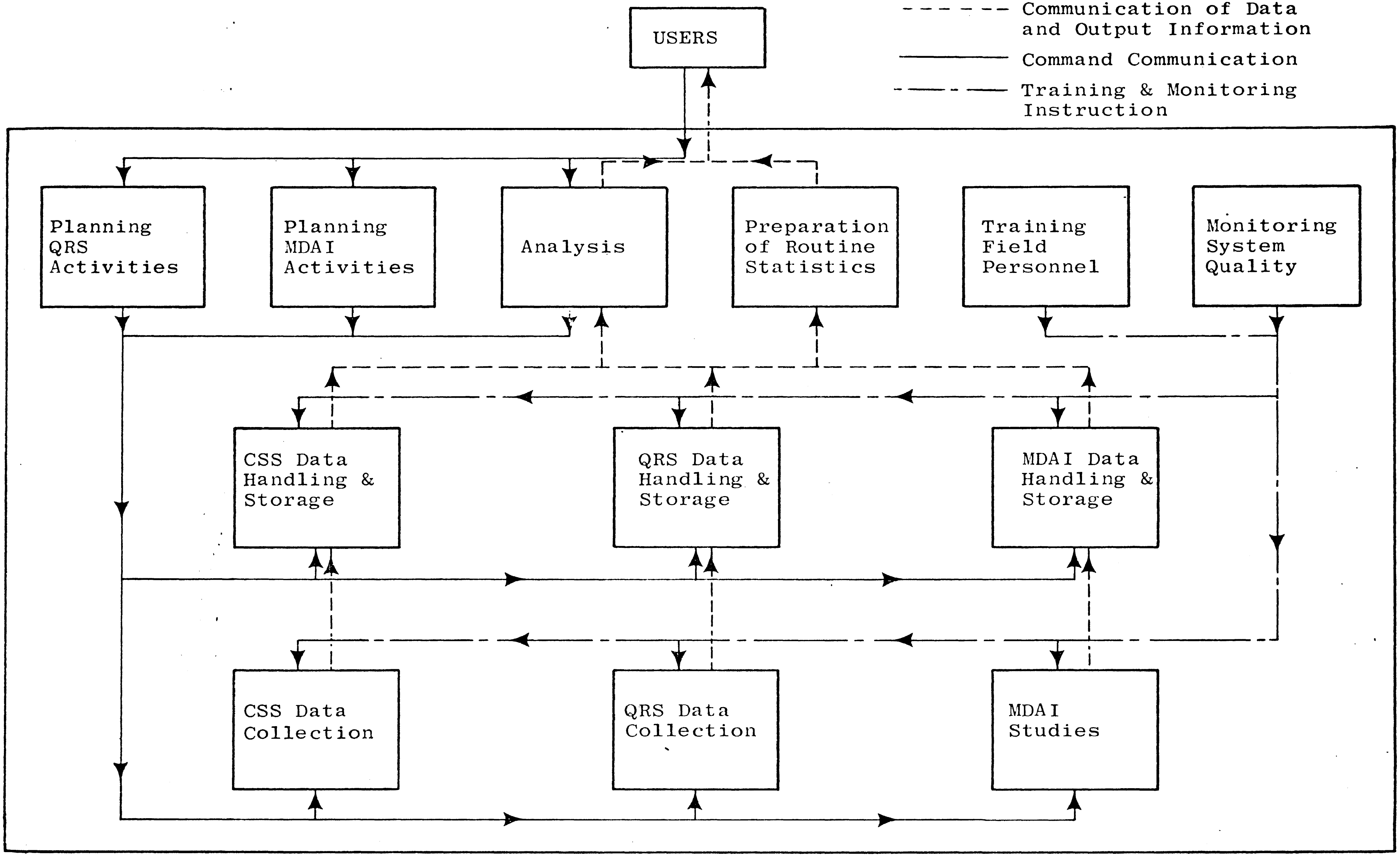

Communication of Data and Output Information
Command Communication
Training & Monitoring Instruction
USERS
Planning QRS Activities
Planning MDAI Activities
Analysis
Preparation of Routine Statistics
Training Field Personnel
Monitoring System Quality
CSS Data Handling & Storage
QRS Data Handling & Storage
MDAI Data Handling & Storage
CSS Data Collection
QRS Data Collection
MDAI Studies

FIGURE 11.1. Functional Block Diagram - National Accident Sampling System

The centralized arrangement is depicted in Figure 11.2. Only the data collection functions are located anywhere other than at the single management center. Data collection organizations are shown in a variety of configurations--one primary sampling unit collecting only CSS data, one only QRS data, etc. But the lines of data flow indicate that the raw data would be transmitted in some way directly to the control center, and commands for change, training, instructions relative to control quality, etc., would go directly from the control center to the field unit. A system similar to this is employed by The University of Michigan's Survey Research Center for its quarterly national surveys, with more than 60 primary sampling units all being managed from a central point. Control of quality and training are also handled centrally, although there is some need for travel and frequent communication before and during a sampling period. This concept was considered for the NASS, but judged inappropriate because of the continuous nature of the data collection and the requirements for occasional if not frequent technical support in the field.

A rather fully decentralized organization is shown in Figure 11.3. In this configuration the headquarters component provides general direction of the program, conducts some analysis (though largely of processed data), and develops general plans for changes in the program. The field units are each shown as incorporating all of the data collection functions, being largely responsible for training and quality control, and prodųcing reports rather than data. The present MDAI system operates largely in this manner, with the principal outputs of the MDAI teams being paper reports forwarded to the management center for

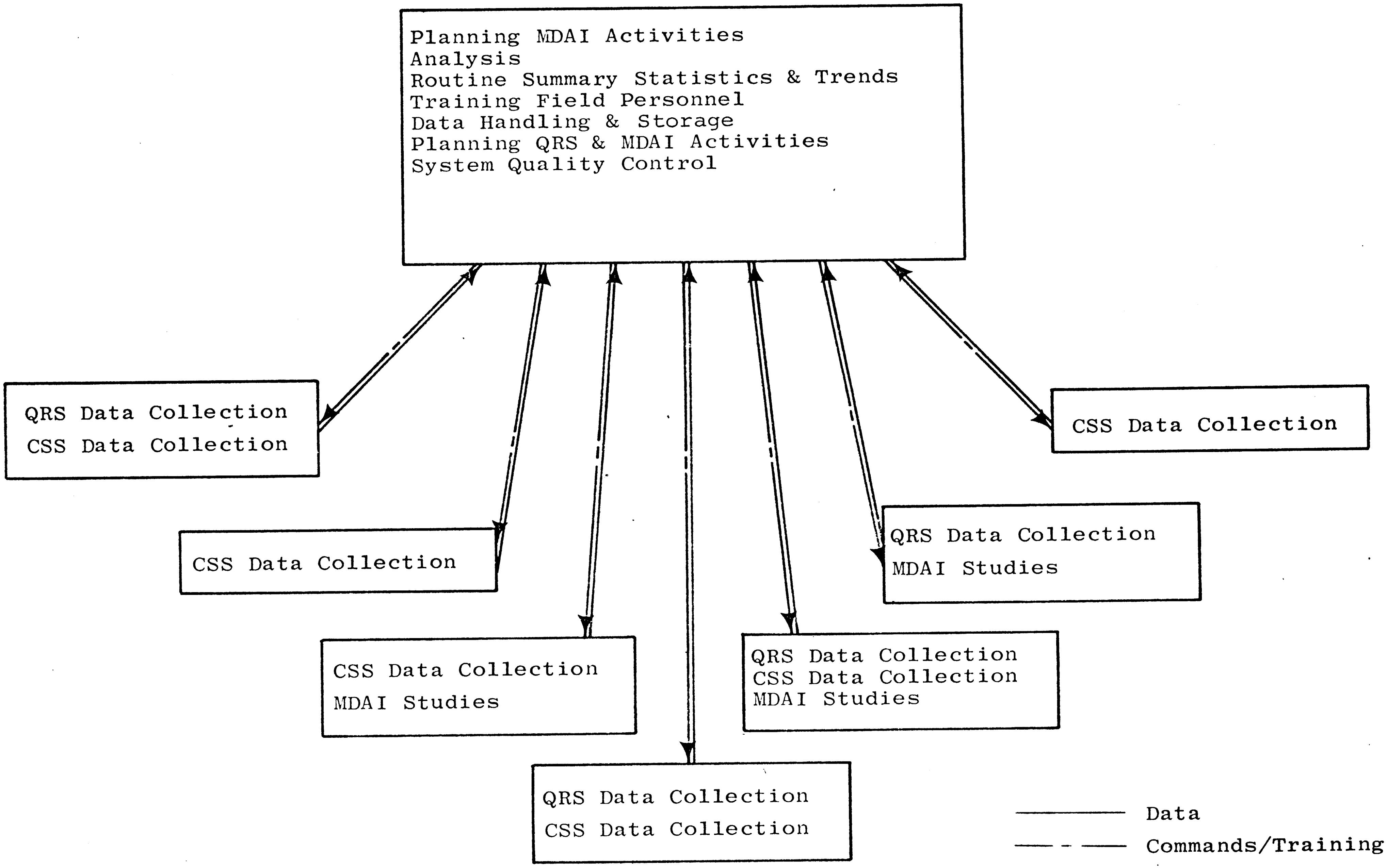


FIGURE 11.2. NASS Centralized Organization

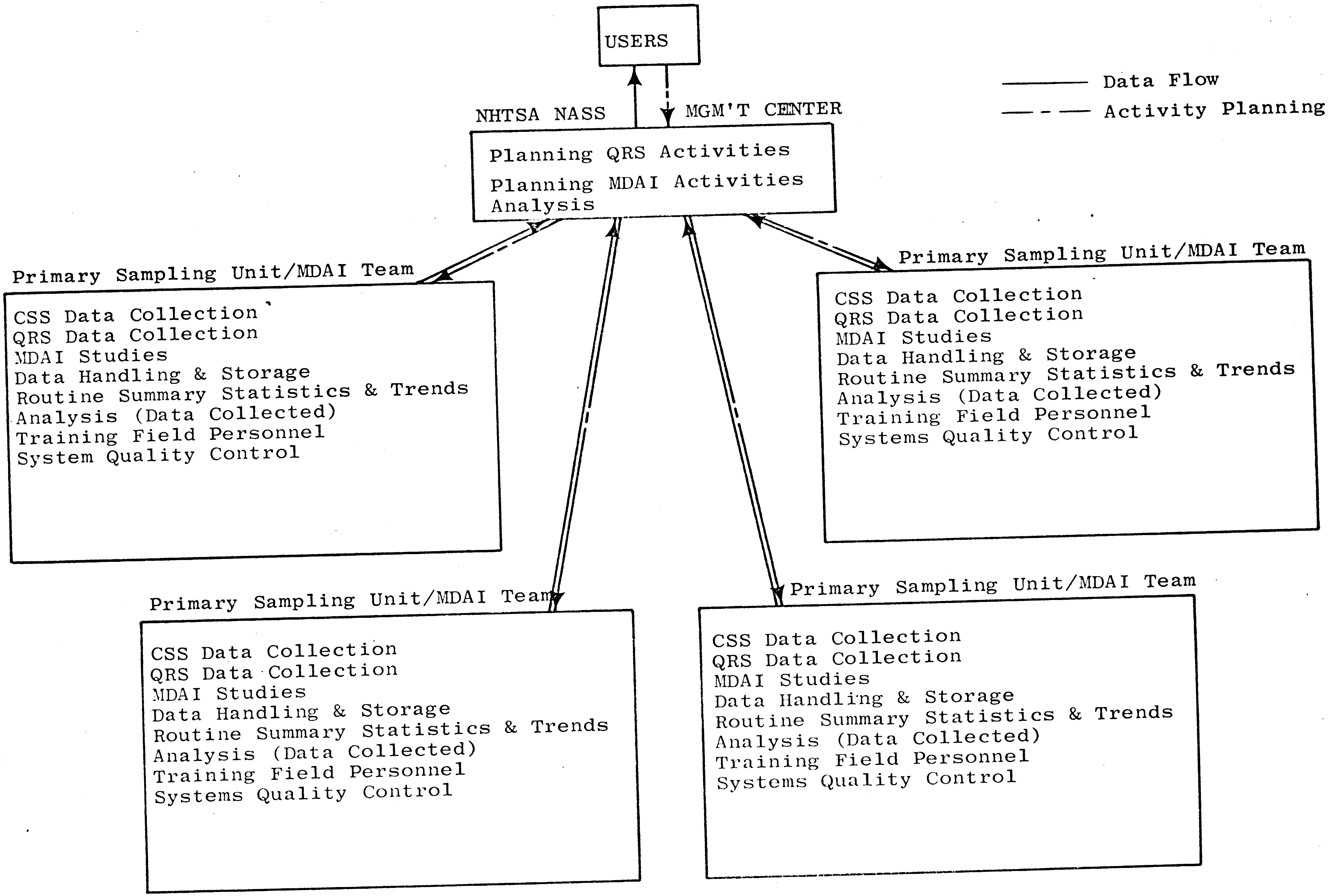


FIGURE 11.3. NASS DECENTRALIZED ORGANIZATION

further analysis. The management center is basically responsible for communication to the world outside the system. This configuration was judged inappropriate for the NASS because of the requirement for consistency of training and the need for rapid reaction to maintain control of quality in the data. Further, in order to develop national statistics, raw data need to be centralized.

The recommended hybrid organization is shown in Figure 11.4. Because of the nature of the MDAI investigations--being detailed studies of a limited number of accidents--only a small number of teams should be assigned that function. The people required for these studies are at a higher professional level than those doing the CSS and QRS data collection, and they could well serve to support CSS and QRS teams on an occasional basis. A part of the management function is distributed to the MDAI teams, which then operate dually as MDAI investigators and as zone managers, each being responsible for five or six primary data collection units.

Command lines proceed downward on the chart through the zone centers. Quality control is still a major function at the NHTSA management center, but the zones have a similar responsibility in monitoring teams under their command. Data, in general, flow directly from the source to the data center, although a digital communications system will permit rather close monitoring of team activities, as described in Section 10 of this report.

Communication with the world outside of the NASS is again largely through the NHTSA management center--both for input and output. Overall guidance of the training activities remains the responsibility of the

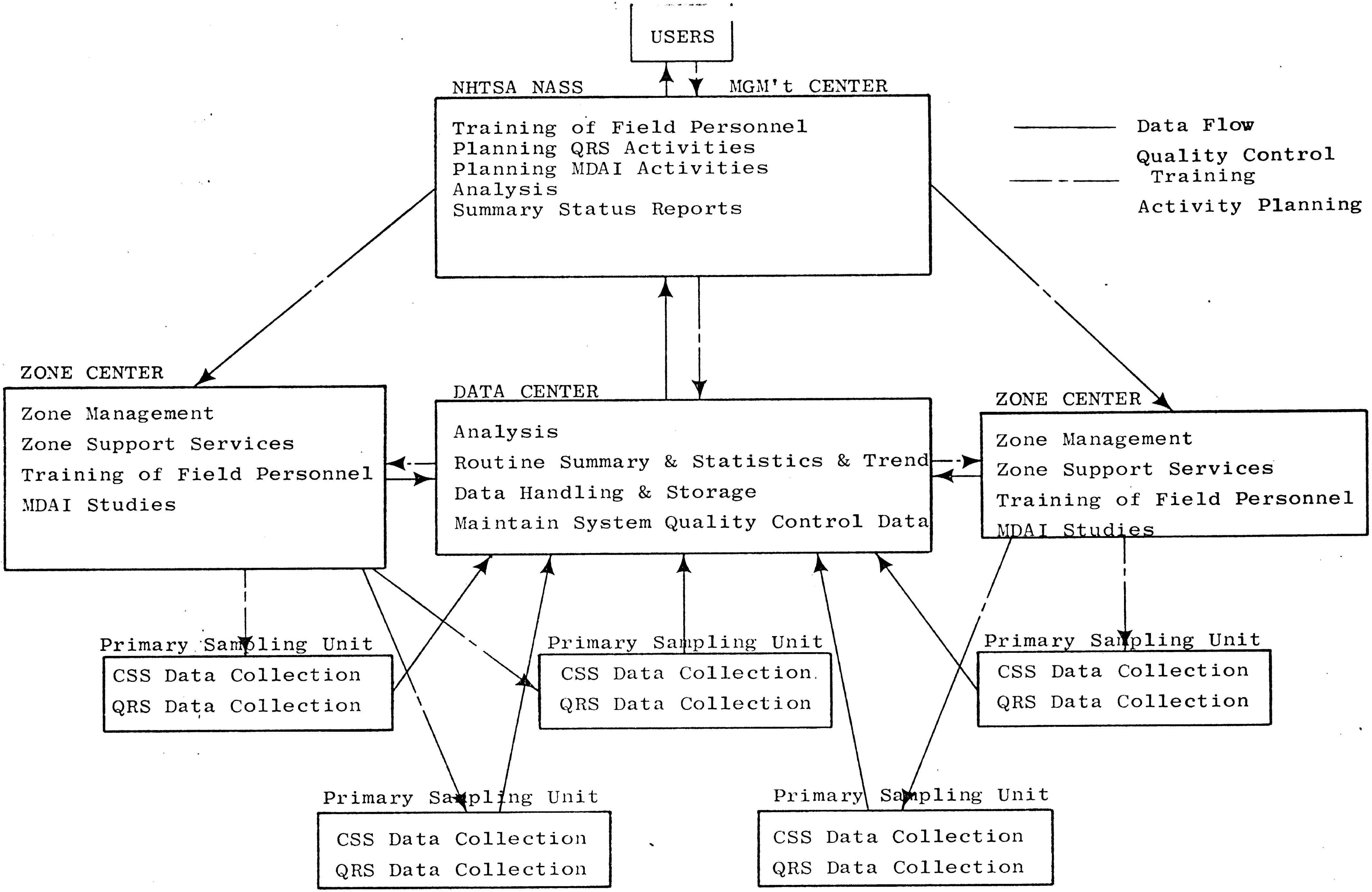


FIGURE 11.4. NASS HYBRID ORGANIZATION

NHTSA center, but the continuing nature of the data collection requires relatively continuous training activity at thé zone level.

Finally, the hybrid arrangement shown fits well as a transition from the present MDAI program, and takes advantage of the expertise currently resident in that system.

# 12. IMPLEMENTATION OF THE NASS PLAN

Although the majority of the design choices for the National Accident Sampling System can be made now, many details can be better specified after gaining more operational experience. The Restraint System Evaluation Study, which took place over the past year, has served a useful purpose in attempting to introduce a careful sampling plan for investigation of a particular problem. The collection of police-reported towaway accidents, the detailed investigation of a well-defined sample of these accidents, and the compilation of the resulting data into a single file, have all been activities which would be conducted in the NASS. Some of the problems in the RSES program have, in fact, dictated design choices in the NASS. For example, variations in interpretation of the Abbreviated Injury Scale among the several teams became apparent as analysis began, and now has been traced to the coding of lacerations at AIS levels 1 and 2. For one team, reported lacerations have been about equally divided between levels 1 and 2, and for another team about 80% of the reported lacerations are at level 1. Although the Injury Scale was intended to permit consistent reporting, the most likely conclusion is that it has been interpreted and used differently at different locations. The monitoring structure of the NASS has been developed largely to intercept such discrepancies as early as possible, and to provide for removing them.

In addition to the RSES operations, the fielding of a single team to collect the data specified for the CSS portion of the NASS would provide further useful inputs to the design. Both of these should be viewed as a "pre-pilot" activity that would permit trying out the basic mechanics of the system. The original intent of the HSRI program was to implement a trial field operation late in 1974, but this was deferred in favor of further development of the national plan. Such a pre-pilot activity is still viewed as mandatory in checking out the mechanics of the system, and is expected to require three to six months of operation by a team working directly with the data center.

## 12.1 Pilot Program

To proceed carefully but quickly toward full operation of a national system, several questions can best be resolved by means of a modest pilot program. The NASS sample design calls for implementation of primary sampling units in diverse regions of the country and in areas of widely differing population densities. Sub-sampling (i.e., secondary sampling as discussed in Section 4 of this report) will be necessary in many of the primary sampling regions, and the method of sub-sampling needs to be tested. The interrelationship between a zone center and a remote primary sampling unit should be tested. The combining of small counties into population units of at least 50,000 was discussed in Section 4; this implies relatively low population densities and a different approach to data collection than would exist in a large city. Pilot testing of a team in such a sparsely populated region would be useful in determining final personnel and travel requirements for these sites.

Site selection from the sampling plan can be accomplished at any time, in a matter of hours. The sampling frame for primary site selection has been defined in Section 5, and the weights for each defined unit have been computed. All that remains is to perform the actual probability sampling--preferably a manual operation--to choose actual locations. In the pilot operation it seems preferable to select actual locations which will be in the final system, and thus no recommendations are made here for specific sites for the pilot test, except to note that Los Angeles, New York, or Chicago (Cook County) could be included with certainty because, with their large populations, they will be self-representing in NASS.

Other sites recommended for pilot operation include one large urban area remote from its associated zone center (this might be, for example, New Orleans, with a zone center at Miami); one rural (multi-county) site remote from its zone center (e.g., several counties in Texas associated with a zone center in San Antonio); one site in the northwestern United States (perhaps associated with an NHTSA regional office as a zone center); and a site in the northeastern U.S., which is a more densely populated rural area.

Several possible procedures for secondary subsampling have been discussed in Section 4. The problem in the larger (i.e., the more populous) PSUs is that there are many more accidents than are required for the sample, and in order to make the sampling process workable and to minimize costs it is desirable to create a local sampling frame from which to choose cases for detailed investigation. It is important that the method chosen minimize any increase in variance of estimates from the system, and some approximation to simple random sampling is desirable. It is quite possible for the method to be different in various primary sampling units--e.g., a rotated sample through all sub-jurisdictions in one place, a wedge sample in another, and a simple random sample in a third. During the pilot phase, several of these mechanisms should be tried in order to provide better guidance for later operations.

Based on past experience in setting up and operating in-depth accident investigation programs it is estimated that approximately 90 days of preparation time will be required before any useful data can be expected from a new site. This is not to say that some data acquisition cannot be begun, but that expectations concerning the data in the first three months should be minimal. The recommended four pilot sites could be implemented at about the same time, and useful data might be expected, then, at about the same time from all of these units.

The initial data collection would be aimed at satisfying the needs of the continuous sampling system, and careful evaluation of the extent of missing data, problems of interpretation of codes, etc., should be accomplished quickly. Training for the pilot programs should be accomplished centrally, using the data forms and the investigators manual (Appendix G ) with instructional material similar to that used in the current MDAI training programs. Frequent (i.e., nearly continuous) monitoring of the field operations by both national center personnel and zone center personnel is desirable during this phase, with periodic meetings to discuss problems and effect modifications.

The Quick Reaction System (QRS) is a most important part of this design. Its purpose is to provide a convenient means to institute a rapid collection of representative data to address some particular subject. The data handling mechanics for the QRS have been developed in parallel with the same features of the CSS, and are adaptable quickly to a new set of questions. During the pilot operation two QRS studies should be conducted--the first, involving prior planning, taking about 60 days for the field data collection activity; and the second, undertaken on short notice, involving about 90 days from start to finish. The latter would test particularly the command communications aspects of the system.

HSRI has submitted several possible topics for QRS operation to NHTSA for consideration early in the program, but comments were delayed during the period of development of a national sampling approach. Candidate topics discussed in July, 1974, included:

(a) The effectiveness of bumper systems in low-speed crashes

(b) Roadside obstacle accidents

(c) The frequency of unlicensed drivers in accidents (and the characteristics of their accidents)

(d) Young children in accidents, particularly with respect to restraint usage

Each of these would be covered only in a limited fashion by the CSS, and specific additional information could be collected through the QRS for a prescribed period on such accidents to provide a national estimate on a specific subject. A sample set of questions for the "young children" topic was provided in the report "Statistical Inference from In-Depth Accident Investigations," and serves as an example of this topic.

characteristics, etc., have become relatively standardized. But a National Accident Sampling System should not be viewed as stationary; there is room for improvement in many of these areas. Vehicle damage reporting using the VDI or CDC classifications, or reporting of vehicle crush and speed, remains somewhat of an art; this system needs a continuing effort at improving the reporting methodology. Over the past five years CALSPAN has developed the SMAC system, Indiana University has developed a reporting methodology for accident causation, The University of Michigan has extended the Abbreviated Injury Scale to an Occupant Injury Classification system, and several people have suggested modification of crash severity recording (e.g., K. Campbell at General Motors, L. Patrick at Wayne State University). Some of these have proceeded to the point where they are now in actual use. The point here is that involvement of professional-level people in this program on a continuing basis (on the MDAI teams at the zone centers) will provide the interested expertise to continue development of reporting methodology. The needs for this work should be recognized by the NHTSA managers of the system, and a long-term schedule should be devised.

Training for the operational system should be similar to that developed in the pilot program, perhaps with one of the MDAI-zone center activities taking primary responsibility. Note again the importance of precise data taking, implying frequent training of personnel and the maturation of the investigators to the level of a professional accident investigator. One of the reasons for a national sampling system is to achieve consistency in report at a level not attainable in volunteer (i.e., police) reporting, and the permanent training and education function is vital to that.

# 13. SYSTEM OPERATING COSTS

The National Accident Sampling System described in this report consists of 35 primary sampling units responsible for the basic accident investigation and reporting, six zone control/MDAI centers which serve a dual function as manager of several PSUs and as a regional MDAI team for in-depth investigations, one national data center which receives, stores, and processes data for the entire system, and one management/ central control center which directs and monitors the operation of the system. These forty units are in general located in different places.

Costs of operating the entire system include expenditures for salaries, office space, travel, communications, and miscellaneous supplies. In order to study alternative system arrangements--for example, variations in total case load, variations in the number of primary sampling units, etc.--basic unit costs have been assigned. These have been developed from knowledge of present accident investigation team operations at both the MDAI level and the more routine Restraint System Evaluation Studies. In general, salaries have been estimated with a burden of 100%, and direct charges (telephone, rent, travel) have been estimated at cost.

There are clearly differences in cost of living, and consequently in the salary needed for a particular job, in different parts of the country. If this NASS is ultimately set up as a civil service activity, it is likely that there will be little difference from team to team; if, on the other hand, the system is set up by contract, it is likely that there will be wider differences in personnel cost. Salaries in this section have been estimated from data in the U.S. Department of Labor's Handbook of Labor Statistics for 1973, comparing jobs believed to be of similar difficulty and responsibility. While there are a few professional accident investigators presently operating in the United States--for insurance companies, for industry, etc.--there is not a large enough group to be identified as such in the Handbook of Labor Statistics. Salary estimates have been revised upward from the data in the 1973 publication to account for inflation.*

## 13.1 Primary Sampling Unit

Manpower and support for operation of primary sampling units has been estimated for annual investigation rates of 250, 500, and 1,000. Although the number of investigations recommended in this report is 500, it was useful in defining an optimum sampling arrangement to consider lower and higher figures, and these have ben been used in Section 4. The staffing of a typical primary sampling unit office for each of these cases is shown in Figure 13.1. Annual salary for each individual is shown on the chart. Average annual cost of operation of a 500 case/year primary sampling unit is $140,000, with a one-time expense of $2,500 for office furnishings. The number of employees shown is expected to be enough to handle the needs of the continuous sampling system, with enough additional time to undertake QRS studies one at a time and to provide relief for vacations, sick leaves, etc. Emergency manpower supplements would be furnished on occasion by the zone center, but the frequency of these events should be low.

*Job description for personnel in the various locations within NASS are detailed in Appendix H.

| 250 CASES/YEAR | | 500 CASES/YEAR | | 1000 CASES/YEAR | |
|---|---|---|---|---|---|
| Team Chief | $17,000 | Team Chief | $17,000 | Team Chief | $17,000 |
| Field Investigator | 12,000 | Field Investigator | 12,000 | Field Investigator | 12,000 |
| Field Investigator | 12,000 | Field Investigator | 12,000 | Field Investigator | 12,000 |
| Field Investigator<br>(Part-Time) | 6,000 | Field Investigator | 12,000 | Field Investigator | 12,000 |
| | | | | Field Investigator | 12,000 |
| | | | | Field Investigator | 12,000 |
| Secretary/Interviewer | 9,000 | Secretary/Interviewer | 9,000 | Secretary/Interviewer | 9,000 |
| Salary Total | 56,000 | Salary Total | 62,000 | Salary Total | 86,000 |
| Vehicles - 3 | 7,200 | Vehicles - 4 | 9,600 | Vehicles - 5 | 12,000 |
| Office, 600 ft.$^2$ | 2,700 | Office, 700 ft.$^2$ | 2,900 | Office, 900 ft.$^2$ | 4,050 |
| Telephone | 900 | Telephone | 1,200 | Telephone | 1,800 |
| Computer Terminal | 1,300 | Computer Terminal | 1,300 | Teletype | 1,300 |
| Office Equipment | 2,000* | Office Equipment | 2,500* | Office Equipment | 3,000* |
| Supplies | 1,000 | Supplies | 1,000 | Supplies | 1,500 |
| Non-Salary Total | 15,100 | Non-Salary Total | 18,500 | Non-Salary Total | 23,650 |
| Annual Total With 100% of Salaries as Personnel Benefits, Indirect Costs, Fees | $125,100 | | $140,000 | | $192,650 |

*One-time charge, not included in annual total operating cost.

FIGURE 13.1. PRIMARY SAMPLING UNIT - COMPOSITION AND COSTS

## 13.2 Zone Center/MDAI Team

Staffing of the zone center/MDAI team is shown in Figure 13.2. The organization is similar to that of the conventional MDAI team, supplemented by an assistant manager who would be responsible for monitoring and working directly with the primary sampling units. The three "specialists" shown could be either full or part-time, depending on the needs, but all three should be represented. Total cost of operation per year is estimated at $200,000.

## 13.3 Data Center Operations

Staffing of the data center is shown in Figure 13.3. The center would receive and place in storage approximately 90,000 "frame" cases per year, 15,000 detailed cases, and 500 MDAI cases. Files for the first two of these would be built approximately monthly, with special fiels built at the end of six- and twelve-month periods. Programming assistance would be required to create and modify analysis programs, or to develop new standard outputs. Computer costs are based on current rates for a University IBM 370/168.

## 13.4 General Control Center

The general control center for the system is viewed as being located at NHTSA and staffed by DOT personnel. Detailed costs are not given here, but the general functions of that staff are shown in Figure 13.4. The senior analyst would work closely with the data center in specifying needs for both routine and special output from the system. The quality control analyst would monitor the standard comparisons made of the data to ensure that reporting was consistent and proper. The

| | |
|---|---|
| Zone Center Manager | $ 25,000 |
| Assistant Manager | 16,000 |
| Human Factor Specialist | 12,000 |
| Vehicle Factor Specialist | 12,000 |
| Environmental Factors Specialist | 12,000 |
| Consultants (medical, metallurgist, etc.) | 5,000 |
| Secretary | 9,000 |
| | 91,000 |
| Personnel Benefits, Indirect Costs, Fees | 91,000 |
| | 182,000 |
| Travel, Supplies, Telephone, etc. | 18,000 |
| Total | $200,000 |

FIGURE 13.2. ZONE CENTER/MDAI TEAM - COMPOSITION AND COSTS

| | |
|---|---|
| Data Center Manager | $ 30,000 |
| Senior Analyst | 24,000 |
| Programmer/Analyst | 16,000 |
| Operator/Junior Operator | 12,000 |
| Operator/Clerk | 11,000 |
| Secretary | 9,000 |
| | 102,000 |
| Personnel Benefits, Indirect Costs, Fees (100%) | 102,000 |
| | 204,000 |
| Supplies, Telephone | 2,000 |
| Computer Costs | |
| Input 90,000 Frame Cases | |
| 15,000 Detailed Cases | 28,000 |
| 500 MDAI Cases | |
| File Build 15 Per Year Frame | 7,500 |
| 15 Per Year Detailed | 7,500 |
| 4 Per Year MDAI | 2,000 |
| Programming | 1,000 |
| Routine Output & Services (e.g., Dictionaries) | 8,000 |
| Total | $260,000 |

FIGURE 13.3. DATA CENTER - COMPOSITION AND COSTS

Program Director

Senior Analyst

Quality Control Analyst

Training Specialist/MDAI Manager

Part-time Vehicle, Human, Environmental Specialist

Quick-Reaction System Manager

FIGURE 13.4. GENERAL CONTROL CENTER - STAFFING

training specialist would develop and monitor the training activities within the system. Vehicle, Human, and Environmental specialists should be available at least on a consulting basis to discuss problems with the other analysts. The QRS manager would be responsible for planning and directing quick-reaction data collection efforts. Overall, the director of the general control center would be responsible to the NHTSA management for ensuring the system's continued useful operation and output.

## 13.5 Summary of Costs

For a 32 PSU system with six zone centers, total cost is given by:

$$35 \times \$140{,}000 + 6 \times \$200{,}000 + \$260{,}000 + G$$

where "G" represents the DOT management group. Neglecting the latter, the total is $6,360,000 per year for a full operating system. As noted above, both costs and manpower estimates have been based on considerable experience, and although great precision in a time of both inflation and unemployment is not possible, they are believed by the present authors to be reasonable. The bases for the costs has been given here so that the reader with different experience may perform his own computations.

Cost of operation of a pilot system can be estimated using the same figures. Data center costs in a pilot operation will be substantially reduced simply because of smaller computer costs--for a four-PSU operation these are estimated at $10,000. Personnel requirements would be reduced by about a factor of two, and total cost of one year of operation of the data center in such a pilot mode is estimated at $130,000.

Cost of operation of the zone centers during pilot operation would also be reduced, since each would have only one primary sampling unit to manage. These are estimated, then, at $100,000. The total cost for a year of operation in the four-PSU pilot mode would be calculated by:

$$4 \times 140,000 \pm 4 \times 100,000 + 130,000 + G$$

and, again neglecting G, the total would be $1,090,000.